Vol. 43, No. 4 July, 1946

Psychological Bulletin

OPINION-ATTITUDE METHODOLOGY

QUINN MCNEMAR*
Stanford University

I—Problems and Issues. II—Attitudes by Scaling Techniques. III—Single Question Opinion Gauging. IV—Administration. V—Statistical Issues. VI—Study of Changes. VII—Correlates and Interrelationships. VIII—Studies of Morale. IX—Concluding Remarks.

I. The Problems and Issues

Over-all View. This paper presents an appraisal, by an outsider, of the techniques and methodologies utilized in studies of opinions and attitudes. First, a brief word as to the meaning of the terms *opinion* and *attitude*. The common element of most definitions of social attitude is that such an attitude is a readiness or tendency to act or react in a certain manner. No one has ever seen an attitude; an attitude, however real to its possessor, is an abstraction the existence of which is inferred either from nonverbal overt behavior, or from verbal or symbolic behavior. The term opinion is frequently defined as the verbal expression of an attitude. This could mean that one can never hold an opinion unless it is expressed, and that in "gauging public opinion" one goes out to gauge something which may be nonexistent prior to the gauging process. Public opinion is defined as the average judgment or consensus of the individuals of a "public" regarding a given issue, institution, or person.

* This critical review was undertaken at the request of a number of social scientists who believed that a critique of the opinion-attitude methodologies should be made, preferably by someone without prior or vested interests. The writer, having reluctantly agreed to attempt the task, hopes that he has singled out for discussion the more significant of the many issues in this expanding area of research. Acknowledgement is due to Paul R. Farnsworth for suggestions from time to time and for a critical reading of the manuscript. Olga W. McNemar has assisted in the rather heavy library work involved in scrutinizing about 800 references, and she has also served as the writer's critic during the preparation of this report. I am indebted to many individuals in governmental agencies with whom I had contact during the two years beginning with the autumn of 1941 and especially to those consulted during the winter of 1944. I am grateful to the Social Science Research Council for a grant which facilitated this appraisal.

Some writers use the terms attitude and opinion as having very similar if not the same meaning; both are interpreted as having to do with a predisposition to action. Thus conceived, one can have an unexpressed opinion; if it is expressed, it may be taken as one's opinion or as evidence regarding one's attitude. Since the terms are so frequently used synonymously, and since what one investigator calls an opinion study might be called an attitude study by another, and vice versa, we shall not attempt herein to adhere to any rules about the noninterchangeability of these words. A useful, albeit somewhat artificial, distinction can be made on the basis of techniques: *The typical attitude study involves a scale or battery of questions for ascertaining attitudes whereas the typical opinion, particularly public opinion, study leans heavily on a single question for a given issue.* Although the two have much in common, there is sufficient difference to permit separate discussion of certain aspects. Section II of this critique will be devoted to "measuring" attitudes by the use of scales, and the single question method will receive attention in the following section.

In brief, the several techniques used for getting at opinions or attitudes boil down to the simple matter of asking people questions about an issue in order to elicit a response which is interpreted as the respondent's opinion about or attitude toward the given issue. The introduction of high sounding definitions does not alter the fundamental fact that practically all attempts at determining opinion and attitude are at the verbal level—the correlation between verbal and overt nonverbal behavior, the latter dependent in part upon the opinion or attitudinal set, is an unknown and is usually left as an unknown by investigators in this field. This is one aspect of the question of validity to which we shall return presently.

The form of the questions asked by interviewers or in schedules varies greatly. Some call for a "yes" or "no" answer; some may require a rating such as "very much," "much," etc; sometimes a list is provided and the respondent is asked to check or rank the items; occasionally a preference between two possible alternatives is called for; quite frequently the question is of the open-end or nondirective type—the respondent is free to phrase his own response; at times the respondent is presented with a statement or series of statements and asked to indicate whether he agrees or disagrees therewith. Still other variations have been used.

Regardless of the question form and how the answers are obtained, the responses are used to categorize the individual as belonging to one of two or more classes. These classes may represent qualitative or near-quantitative differences. The category to which an individual's opinion

is assigned may correspond exactly to his response or responses; it may depend upon a rating by the interviewer or by the analyst who studies and codes the verbatim interview; or it may depend upon either an arbitrary or rational scoring of the responses to several items. This last procedure is an attempt to assign to an individual some position or rank on a continuum or near-continuum so as to characterize his belief in, agreement with, or feeling about the given issue. What questions or items should be used for this purpose and how they should be weighted is the problem of "scaling" to be treated in Section II.

The studies, which are often characterized as surveys, are always done on a sampling basis. The aim is to secure a sample which, within the limits of random or chance errors, can be thought of as a representative cross section of some defined population. Thus the inferences made or generalizations deduced hold only for the population sampled, and any inadequacies in the sampling procedure should, perforce, impose qualifications on the conclusions. Most of the so-called attitude studies have been made on college student groups, or others which happen to be available. Usually the investigator fails to specify how he drew his sample, and from what population. Sampling issues will be briefly discussed in a later section.

The statistical analysis of the data of opinion and attitude research may be simple or relatively complex: simple if the aim is merely to report an over-all percentage; more complex if the aim is to make group comparisons, to correlate opinions with other variables, or to study the interrelatedness of opinions. The problems of analysis are not much different from those in other fields of social science. Pitfalls and hazards exist and, as will be seen in later pages, some errors and inadequacies in statistical treatment occur rather frequently.

The foregoing sketch does not include all the methodological issues which are to be treated in subsequent pages. There are certain questions concerning the securing of the data—by direct interview, by mailed questionnaire, by group administration of schedules, etc. The study of opinion-attitude changes is of sufficient importance to deserve somewhat detailed exposition. The same is true of the determination of correlates and intercorrelates of attitudes. Some other persistent methodological problems are illustrated in Section VIII.

Fields of Application. The reader who is not already familiar with the rather extensive use of attitude-opinion techniques may find a few paragraphs thereon of interest. It is not enough to say that these methods have been used by journalists, sociologists, political scientists, psychologists, educators, consumer research specialists, *et al.* What we wish to convey here is some notion of the diversity of more specific

fields of application. It will become quite obvious that the boundaries of disciplines are immaterial. It is difficult to say to which group belongs the major credit for developing the techniques, but it is reasonably certain that the psychologists and sociologists are mainly responsible for the current status of attitude research (by scales), and that the journalists have contributed much to opinion research (by single questions).

All readers are familiar with the Gallup and *Fortune* polls and the types of issues—political, economic, international, etc.—dealt with primarily as a service to journalism. The main aim of these polls is the reporting of opinion on various timely topics, and the determination of shifts of opinion regarding certain issues. Usually over-all percentages or percentages for a few breakdowns are placed on record. The results of these surveys generally appear only in newspapers and magazines, seldom in scientific and scholarly journals except for the periodic appearance in the *Public Opinion Quarterly* of current opinion surveys carried out by *Fortune*, the American Institute of Public Opinion (Gallup), and the National Opinion Research Center (Denver).

In the professional journals hundreds of articles on attitudes (mainly) and opinions have appeared during the last fifteen years, with additional hundreds prior to this period. In general, these more remote papers are of less importance than those following the impetus given research in this field by L. L. Thurstone about 18 years ago. By Thurstone or under his direction scales were constructed for "measuring" attitudes toward the church, treatment of criminals, capital punishment, God, communism, patriotism, the Negro, birth control, the Constitution, the Bible, Sunday observance, war, evolution, censorship, the Chinese, the Germans, etc. A few years later a series of scales appeared under the authorship of H. H. Remmers and his students. These were so-called generalized scales designed to measure attitudes toward any social institution, any school subject, any teacher, any dramatic production, any disciplinary procedure, any practice, any racial or national group, any proposed social action, any selection of poetry, any vocation or occupation, any homemaking activity, in fact nearly *any* anything. In order to pave the way for grinding out more scales of this type, one graduate student (**100**, pp. 117–126) was set, or found, the task of building up lists of the multitudinous miscellany about which people can and maybe do have attitudes.

Attitudes toward rural and urban life, the social rights of Negroes, many economic policies, the T.V.A., relief, and the public press have been studied. Attitudes regarding law, juvenile delinquency, behavior problems, and various aspects of religion have been investigated. The attitudes of children toward parents and of parents toward children have also been tackled. Education, nursery schools, the high school, and intercollegiate athletics have been the object of attitude research. Opinions about radio advertising, the movies, retail stores, women executives, feminism, the Dies Committee, political parties, marriage and sex have been collected and analyzed. Many publications have been concerned with that complex of attitudes and opinions having to do with religious, political, social, and economic liberalism-conservatism.

The various techniques have been utilized for the study of job satisfaction, employee attitudes, several aspects of the morale of civilians and soldiers, optimism-pessimism, and satisfaction with life. Attitudes toward the future and toward death have not escaped notice.

It can be said that some of the questionnaires used in these surveys contain everything but the proverbial kitchen sink, and once such a questionnaire has been filled in by a sizable group its author has the "basic" data at hand for a half dozen articles. If he is fortunate enough to have punched card equipment, it becomes the misfortune of his professional contemporaries to find the literature being filled with results of cross tabulations which are so lacking in rationale as to be nonsensical. The "hypothesis" step in scientific reasoning and research seems to be all too frequently ignored by the users of these techniques. If by 1931 a whole book (132) could be devoted to the attitude field, one might expect that from then on enough facts about attitudes would have been known to have permitted systematization sufficient for fruitful theories and hypotheses.

Some Fundamentals. It may be advantageous next to discuss briefly a few of the fundamental requirements which all techniques for "measuring" attitudes and opinions must meet. Certain of these are so basic that unless reasonably well attained, it cannot be expected that the study of attitudes and opinions will ever approach scientific status. As will be seen presently, the attitude scale or battery of questions and the opinion or single question methods differ somewhat in the extent to which an attempt is made to meet these basic requirements. If one is to judge from current practices, it can be inferred that some attitude-opinion researchers will brush these requisites aside as being old-fashioned.

The presumption, of course, is that attitudes and opinions are being measured or gauged by use of these methods. The extent to which the terms "measured" and "gauged" are applicable, or just what these terms can connote, needs to be scrutinized. If one is attempting to measure or gauge something, the accuracy or error of measurement must be known. This problem of reliability is common to the physical, the biological, and the social sciences. A very large number of the variables of science are measured indirectly, and such indirect measurement raises the question of validity—the extent to which the instrument is measuring the variable it is designed to measure rather than reflecting some other variable or variables. It is nearly always desirable that a given instrument should measure just one dimension, i.e., involve one continuum.

Measurement. What can this term mean in opinion-attitude research? A number of critics of psychological measurement in general, and attitude measurement in particular, have rightly claimed that such is not measurement in the true sense because nothing is known concerning the equality of the units used in the scales. At first thought, this

criticism might seem quite disturbing, but actually it only means that certain limitations of these so-called scales must be kept in mind. If we have A scoring 4, B scoring 6, and C scoring 8, it simply cannot be said with any certainty that A and B differ as much as C and B or that C possesses twice as much of the attitude as A. What can be said is that B's value differs from that of C in the same direction that A's value differs from B's, a statement which assumes that a single continuum is involved. In order to be sure that C is more favorable in his attitude than B who in turn is more favorable than A, it would be necessary to know the possible magnitude of the "error of measurement."

It should be understood from the outset that no claim is made herein to the effect that attitude measurement permits more than a rank ordering of individuals. The single question or opinion gauging technique does not accomplish even this much, but it does permit a rank ordering of groups of individuals.

Reliability. In discussing the concept of reliability it is necessary to distinguish between two current usages of the term. Its more general meaning seems to be prevalent among the opinion pollers, whereas a restricted meaning seems to have been accepted by the attitude testers. In its more general sense it means the dependability of a result—thus encompassing not only the restricted, or accuracy, idea but also the notion of validity and sampling adequacy. By reliability we shall mean the accuracy with which an individual's attitude is measured, or the degree of error involved in assigning an individual to a class or in establishing his rank-order position. This is the commonly accepted usage among psychologists. Reliability in the sense of sampling dependability will be subsumed under sampling stability of statistical measures. The term validity will also be used in a restricted sense: Does the test, scale, or question tap the variable it is supposed to measure? Thus we shall *not* speak of the reliability of a percentage or of the validity of a study.

All measurement is befuddled with error. About this the scientist can and does do something; he ascertains the possible extent of the error, determines whether it is constant (biasing) or variable, or both, and ever strives to improve his instruments and techniques. Physical scientists usually determine the magnitude of measurement errors by making several measures on the same object, then expressing the error as the average error (in terms of original units) or as the percentage error (the average error relative to the object's magnitude).

In the psychological measurement of a trait or characteristic of a person, it is impractical to make more than one or two determinations for a single individual, hence the amount of error cannot be found in the same manner as in the physics laboratory. This apparent predicament is overcome by taking two measures on a large number of individuals, say 100 or more. Although the average discrepancy could be

taken as indicating the error, it has been found more convenient to compute the product moment correlation coefficient between the two sets of measures, which sets are assumed to be comparable measures of the same characteristic. It is well known that from this reliability coefficient and the standard deviation for the trait as measured, one can state the "standard error of measurement," which is analogous to the "average error" in that it is in the units of the test or scale. Relative error in the physicist's sense is inapplicable here because of the lack of a real zero point, but a relative measure is available in the reliability coefficient itself which is inversely proportional to the error variance *relative* to the trait variance.

It is, of course, well known that the nearer the reliability coefficient is to unity, the more accurate the measurement. Since perfect reliability is impossible, and particularly difficult even to approach in psychological measurement, the question is frequently raised as to how high the reliability must be for a given purpose. To this there is no answer which will be satisfying to all. The writer is of the opinion that, if a study involving measurement is worth doing, every effort should be made to attain accurate measurements rather than being satisfied with just any reliability. The writer is also of the opinion that the 18-year-old dictum of one of our better known statisticians to the effect that a reliability coefficient of .50 is sufficient for group comparisons has been a disservice to the cause of adequate measurement. Far too many investigators have hidden their heads in the quicksand of that dictum. One of the characteristics of mediocre research is mediocre reliability of the instruments used.

The specific reliabilities attained by attitude scale techniques will be given in the next section and what little is known of the reliability of the single question method will be presented in the third section. Suffice it to say here regarding the latter that the accuracy with which one is classified as, for example, favorable or unfavorable to a given proposition can be ascertained even though no expression similar to error of measurement is possible. A type of correlation can be calculated providing two responses are obtainable from each of several individuals. These two responses can be to two different questions designed to get at the same opinion; or a second response to one question can be secured at a later date. This is not the place to discuss which of these two procedures is the better, or to elaborate on the difficulties involved in either scheme. It may be noted, however, that some have objected to the second scheme because it is feared that bona fide changes may have taken place; thus an individual would not tend to give the same answer twice. This poses a dilemma: If an opinion or attitude is such a momentary thing that individual reliability or consistency is lacking, how can one expect the opinion in question to correlate with stable sociological or psychological characteristics of the individual? The research problems associated with unstable as opposed to relatively stable opin-

ions and attitudes must be quite different, yet we fail to find this distinction in the literature.

The factors which lead to variable errors are well known, and need not be catalogued here. The nature of possible constant errors as an aspect of individual reliability differs somewhat for the attitude scale and the opinion question. If a neutral position on some issue is of importance per se or as a point of reference, then the selection and weighting of the questions for a scale and the wording of the single opinion question have an important bearing on biasing the score or the response. But in case of an attitude scale having no absolute neutral point, it is difficult to see what bias could mean. A constant error of four points would only result in shifting the average, which average has no absolute value anyway. For the single question with or without a provision for a neutral response, bias can of course be introduced, perhaps never avoided, but such bias would not act as a constant error since not all individuals are affected. The biasing error definitely affects some, and therefore the reported over-all percentage for a group will be biased. Variable errors, however, would not lead to a predictable error in a percentage, but would add to the variable error already inherent in the percentage as a result of sampling.

Validity. The problem is essentially that of supplying evidence that the device used measures or classifies the attitude or opinion it was designed to measure. This is an old problem which has engendered much debate, and from which many avenues of escape have been attempted. Does an expressed opinion actually represent a person's real attitude? Can verbal or symbolic behavior be depended upon as indicating an individual's action tendency? What is the degree of relationship between overt nonverbal and verbal behavior? The problem is sharpened by reference to the work of La Piere (70) who found that practice and verbal behavior on a race issue were markedly inconsistent.

Validity can be ascertained by a variety of methods, which will be listed here without much comment. Sometimes verbal behavior can be checked directly against nonverbal overt behavior, and a correlation thereby established. Evidence for validity can be obtained by comparing those of known attitude as inferred from their activity or voluntary membership in organizations. Presumptive evidence can be secured by learning whether the scale or question differentiates between groups which on a priori grounds should differ in their opinion about the given issue. Sometimes it is possible to utilize the ratings of close acquaintances as a criterion for checking validity. All too frequently a new scale is checked against an older scale, the assumption being that a high correlation between the two indicates validity. This practice has been overworked by the mental, the attitude, and the personality testers who are seldom too careful about first being sure that the original scale's validity has been established. Another scheme which might aid in ascertaining validity would be to follow the scale administration or the

single question by intensive and extensive interviews in order to see whether the first expressed position holds up under cross examination.

It should be noted that any of these schemes for getting at validity can be followed out on a miniature cross section of 100 or more cases. All of these schemes involve information external to the scale itself. Opposed to the use of external criteria is the a priori method of validating. A scale for attitude toward an issue is said to be valid because the items in the scale make it so, that is, the items have been judged valid. This does not overcome the fact that the verbally administered scale calls for verbal responses which may or may not reflect the individual's attitudinal set. The same is true when an opinion is elicited as a response to a single question. As an illustration, let us look at some results reported by Corey (27) on attitude toward cheating. A scale to measure this attitude was constructed by one of the standard procedures. The reliability coefficient of .91 is adequate. When scale scores were correlated with actual cheating behavior, the relationship was an insignificant correlation of .02. Evidently the scale did not tap the action tendency for which it was designed. One cannot, of course, always be sure that attitudes can be correctly inferred from overt behavior—it is possible for such to be an act, or conforming behavior. A study by Schanck (114) has led to an attempt to distinguish between privately held and publicly expressed opinions. For example, at church a man will publicly take a stand against card playing although he privately admits that he favors card games.

Some investigators have sidestepped the problem of validity by denying that anything exists beyond the verbal expressions, hence there is no problem of validity. Others have adopted the idea that scales or questions test whatever they test, so why worry.

A word should be said about item validation via internal consistency. It is the writer's opinion that internal consistency has to do with reliability and that validity cannot be inferred therefrom.

There are several obvious factors which affect validity and/or reliability. It is difficult to see how a reliable and valid response can be secured unless the respondent understands the given issue. Likewise, he must be able to understand the questions asked him. Ambiguous wording must be avoided. Other things being equal, the less the personal relevance of an issue, the lower the reliability and validity. Stereotypes and emotionally charged words or phrases, though possibly leading to greater consistency, are not conducive to high validity and reliability. Some of the disturbing factors in the wording of statements and questions for attitude scales can be minimized by following the rules set forth by Wang (125). The problems involved in the wording of questions for opinion polling will receive attention later, as will also the influence of interviewing as contributing to possible erroneous responses.

In connection with the subject of validity, it might be noted that the absence of any objective evidence for the validity of a new scale

does not dampen the tendency to give it a name. Once named, its author and other users all too frequently find it convenient to forget that its validity is unknown. A still worse practice is the tendency of some to rename a scale which was originally constructed as a measure of a particular attitude, and thereafter treat it as a measure of some other attitude. For example, scales constructed (validated?) by the Thurstone technique to measure favorableness of attitude toward God, the church, and war, have been interpreted as measuring liberalism-conservatism (63). Perhaps this is true but where is the evidence, and what does liberalism toward God mean?

Single continuum. Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank order are quantitatively and, within limits, qualitatively similar in their attitude toward the given issue. As an example, suppose a test of liberalism consists of two general sorts of items, one concerned with economic, the other with religious issues. Two individuals could thus arrive at the same numerical score by two quite different routes. Now it may be true that economic and religious liberalism are correlated but, unless highly correlated, the meaning of scores based on such a composite is questionable.

This leads to the problem of general versus specific attitudes, a controversy which can best be settled by resort to further investigation. At this place we are only trying to point out that one fundamental requirement for an attitude scale is that it place individuals in rank order along a single continuum. The intelligence testers have gone through a cycle: first the measurement of simple functions, then of more complex and general functions by a hodge-podge of items, and recently, partly as a result of factor analysis, the stress is on the development of tests, each of which involves a single continuum or a so-called unitary ability. Attitude testers have gone through the first two phases of this cycle, but some are still willing to tolerate a morale attitude score made up of four uncorrelated components! This haphazard adding of dissimilar parts on the assumption that a meaningful whole will result is nonsense.

Upon first thought it might seem that this single continuum requirement need not be considered when individuals are to be classified by the single question method. It is true that no continuum is apparent, unless one is inferred as underlying the expressed opinion, but a similar requirement does exist in that it is hoped that those who are classified as favoring an issue are relatively homogeneous. To be classed together should mean similarity, and this can be attained, at least approached, by the use of a question which is unambiguous, which imposes a similar reference point on all respondents, which is so easily understood that personal interpretation as to the meaning of the question is ruled out, and which leads to an unqualified answer.

The single dimension requisite is not only pertinent when a score or response is being interpreted; it has an important bearing on any research which is designed to analyze individual differences in attitude, i.e., to determine the correlates of attitudinal differences. Such an attempt at "explaining" variation is handicapped if the variability under analysis is rendered more complex by the presence of many continua in the measuring instrument. One does not, of course, get rid of all the complexities of social psychology by isolating unitary characteristics and developing "pure" tests thereof. This is only a step in the direction of avoiding the one complexity introduced by failure to construct scales for measuring one dimension at a time.

Another aspect of the single continuum problem is the meaning of the two extremes on a scale. If one end indicates a favorable attitude, does the other end represent unfavorableness or something else? For instance, is Fascism the opposite of communism? Such would seem to be the case in the minds of some. Are we dealing here with one or two dimensions? To interpret the extremes as being real opposites or to interpret the extremes as on different continua is precarious in the absence of knowledge concerning the real meaning of the extremes.

These basic considerations regarding units of measurement, reliability, validity, and dimensionality will again come to our attention at appropriate places during the following detailed discussion of attitude and opinion measuring techniques.

II. Attitudes by Scaling Techniques

Pursuant to our rather arbitrary distinction between attitude measurement and opinion gauging, we shall now consider the methodological problems involved in the construction of scales which permit rank ordering of individuals into a number of graduated classifications. For convenience, scaling methods are being separated from the single question or polling method.

The aim of scaling is the development or construction of a "measuring" device which will distribute individuals along a continuum running from a highly unfavorable through neutral to a highly favorable attitude. The number of steps or gradations, sometimes called discriminability, which is required will depend upon the fineness with which one wishes to differentiate between individuals. Other things being equal one can assume that the finer the possible gradations, the more reliable the instrument, but if one can secure satisfactory reliability with, say, 50 possible scores or positions, there would seem to be no reason for seeking greater discriminability. In so far as the likely statistical treatment of results is concerned, it would seem that from 12 or 14 to 20 possible (and actually found) gradations should be sufficient. This follows from the well-known fact that statistical terms such as

means, standard deviations, and product moment correlation coefficients are not seriously in error when computed from coded values providing that at least 12 or 14 code values are used, and the gain from having more than 20 is negligible. Additional possibilities should be considered. In the first place, the fact that, say, 16 steps are available does not guarantee that the scores will fall over a range of 16 values and, secondly, even if the possible range is utilized in one group, the instrument may not permit fine enough classification when groups of greater homogeneity for the given attitude are being studied.

The assumption that individual differences in a given attitude can be thought of as falling on a continuum has brought forth vigorous objections. It is said that this may or may not be the case. Suppose that it has been demonstrated that a certain attitude does not lie on a single continuum, it does not follow that measurement is impossible. Instead of trying to place individuals along just one continuum, it becomes necessary to develop scales for as many continua as have been demonstrated to exist. How many is an empirical problem, as is also the related question of the extent to which attitudes are general as opposed to specific. The burden of proof as to whether an attitude as measured does follow a single dimension falls upon the one who constructs the measuring scale, or develops the scale construction technique. The requirement that a scale shall reflect a unitary attitude does not rule out the possibility of attempting to measure general attitudes or complex attitudes or specific attitudes. The complex ones would need to be broken down into component parts and a scale provided for each part. It may be that some so-called attitudes will not be amenable to scaling as unitary characteristics.

The early efforts at attitude or opinion measuring usually involved a questionnaire or battery of questions which were selected on an a priori basis. Numerical values were assigned arbitrarily to the item or question responses and these values were summed to secure scores which were then interpreted as indicating the attitude of the respondents. There was nothing about the procedure to guarantee that any one item tapped the same attitude as the other items. Little thought was given to the number of dimensions—that more than one was frequently involved seems obvious. Reliabiiity was occasionally determined, but validity was seldom mentioned.

An attempt to improve on this haphazard work was reported in 1925 by F. H. Allport and Hartman (1) who reasoned that the questions or statements to which responses were to be secured should be so chosen as to represent varying degrees of favorableness or unfavorableness. Their report came to the attention of L. L. Thurstone, who saw that the assignment of statements to a favorable-unfavorable continuum was

a problem amenable to solution by various psychophysical methods, the most feasible of which was the method of equal appearing intervals.

A detailed discussion of the Thurstone attitude scaling technique need not be given here; any reader who is unfamiliar with the method can turn to a number of sources, particularly a 1928 paper by Thurstone (**122**) and the monograph by Thurstone and Chave (**123**). Suffice it to say here that the scheme is one for arranging items on an 11-point scale according to the degree of favorableness or unfavorableness as determined by having a large number of judges sort quite a number of statements into 11 piles. The median judged location for an item is its assigned scale value. Objective criteria were developed for eliminating items which are ambiguous and which are irrelevant in the sense of not belonging to the continuum from the viewpoint of the judges as sorters. The process, of course, begins with the assembling of a much larger number of statements than needed in order to be able to select, from those not eliminated on the basis of ambiguity and irrelevancy, a desired number, say 20, with scale values nearly evenly spaced along the assumed continuum. With sufficient statements, two comparable forms can readily be constructed.

In taking the test, the respondent is instructed to check those statements with which he agrees, and his score is the median of the scale values of the checked statements. Such scoring permits a very large number of score positions. The reliability coefficients computed by the form versus form method for scales so constructed, for a variety of attitudes, are usually between .70 and .90, with typical values in the low .80's.

That the scales may possess validity has been demonstrated by using groups of known (?) attitudes, but just how high or low the validities are has not been adequately determined.

There have been numerous and varied criticisms of the Thurstone method, many of which would have been forestalled by a careful reading of Thurstone's publications. All attempts to disprove his working assumption that the determined scale values will be independent of the attitude of the sorters toward the given issue have been futile. The claim that higher reliability would result if the mean, instead of median, scale value of the checked items were used in scoring seems unsubstantiated by research on this problem by Lorge (**81**).

Some have been critical of the units along the base line or continuum, claiming that these units are not equal. The present writer has been unable to find that Thurstone ever claimed equal units for his scales—he has spoken of equal *appearing* intervals. That any scheme utilizing statistical measures based on a sample can ever lead to equal units would seem impossible because of the presence of sampling errors. Thurstone has claimed that scales constructed by his method for different attitudes would permit the direct comparing of scores as between two attitudes. To this there can be no objection providing a strict comparability of

scores is not assumed. The writer has not been able to see that the defined neutral points for two different attitudes are really the same. If the average score for the attitude of a group toward capital punishment is 6.5 and toward law is also 6.5, can it be said that the group averages as far from a neutral position on one issue as on the other? The type of comparability of scale values which has been rather generally accepted in other fields of psychological testing is that percentile ranks, or standard scores for normalized distributions, denote similar positions. If such is the case, then scales by the Thurstone technique definitely do not yield comparable units; if the Thurstone method does lead to similar metrics when applied to different attitudes, then the intelligence testers err in assuming the comparability of percentile ranks. Attitude and aptitude variables are, of course, quite dissimilar but when "measured" one notes the common element: individuals are assigned scores which are supposedly indicative of the degree of ability in one case and the degree of favorableness in the other.

It is the writer's opinion that the comparability of scales constructed by the method of equal appearing intervals has not been sufficiently well demonstrated to permit the direct comparison of means based on different scales. Remmers and followers (**99, 100, 101**) have been the most persistent in this practice.

A recent study by Riker (**104**) may have some implications for questions regarding the neutral point, and the comparability of values. He used six scales which had been constructed by Thurstone, or under his direction, and 218 Princeton undergraduates who were asked to check their degree of favorableness-unfavorableness on an 11-point graphic self-rating scale and also their intensity of feeling on an 11-point self-rating scale. Thus Riker had distributions for Thurstone scaled scores, and for two types of self-ratings for attitude toward the Negro, the Germans, the treatment of criminals, capital punishment, evolution, and communism. He reported that the largest difference between means for Thurstone versus rating scale values is .916 times its standard error, hence the differences could easily be chance. He concluded from this that the three methods yield "similar metrics." If this were so, we would have some confidence that the neutral points for the three are comparable, but it turns out that Riker made some kind of a consistent error in determining standard errors. Our own calculations, with which corrections recently published by Riker (**105**) agree, indicate that the critical ratios range from .7 to 19.4. Enough (15 of 18) of these critical ratios are sufficiently high to lead us to conclude beyond any doubt that these three schemes do not yield "similar metrics," a fact which also should have been apparent to Riker had he regarded the failure of the three methods to yield similar standard deviations.

It follows, therefore, that a group with mean scale score at the defined neutral point will not necessarily rate themselves as neutral in attitude. It may be that the neutral point on a scale constructed by

the Thurstone technique does correspond to "real" neutrality on a scale of absolutes, but this has not been definitely proven. The writer would grant that the presumptive evidence favoring the assumption is far greater for the Thurstone type of scale than for the less rational scales for which the neutral points are arbitrarily designated as the mid-point of the possible scoring range.

Kirkpatrick and Stone (68) have criticized the Thurstone method on five counts, which actually boil down to three: inequality of units, questionable assumption of attitude continua, and ambiguity due to different motives or reasons back of responses to different items. The matter of continua has already been discussed herein. As to inequality of units, the criticism is that the interchangeability of units on the base line has not been demonstrated. The same criticism has been made more forcefully by others, particularly by Johnson (62) and Merton (86). With the criticism we agree (see Section I), but we question whether any progress is made in Kirkpatrick and Stone's solution to the problem. They proposed that items which had been retained on the basis of the judges' agreement (1) on whether the item was favorable or unfavorable, (2) on the point of reference as social, personal, cosmological, or epistomological, and (3) as to "vigor of wording," should be simply scored as plus one and minus one for pro and con respectively. They reasoned that "At least in the method of scoring used there is a counting process and a very simple manipulation of multiples of units which are relatively interchangeable," and hence concluded that a quantitative variable had been derived. They have not shown that any two statements lead to responses which are even relatively equal so as to make the score of "plus one" on one item equivalent to a "plus one" on another. Counting apples is not measurement even though very simple manipulation is possible. It would seem that the real meaning of interchangeability of measurement units has not been understood by these authors. Elsewhere Kirkpatrick (67) says that inter-person interchangeability is lacking. Is that what the physical scientists mean by interchangeability of units?

It is also of interest to note how Kirkpatrick and Stone handled the problem of ambiguity of reference point. As indicated above, items were not retained unless there was agreement as to the basis of argument or point of reference. As we understand their report, they did retain items which involved at least four different reference points, then proceeded as though they had little faith in their own valid criticisms regarding the use of questions which may be responded to from different viewpoints.

It is the conviction of Merton (86) that since Thurstone's method does not lead to units which are additive and interchangeable on a linear continuum, the scales should be abandoned in favor of item analysis, which he considers "one way in which opinionnaire results can be legitimately employed without recourse to dubiously applicable mathe-

matical operations." He reasons that for a true scale, persons endorsing an item graded as very unfavorable should endorse all statements which are slightly less unfavorable. Ideally, yes, but Merton who says that the advocates of scaling are too far from the "facts of sociology" ignores the just as well established psychological fact that the responses of individuals to nearly all types of questions are subject to response errors. These errors not only complicate the problems of true scaling but they also disrupt analyses based on single items; the latter fact is not mentioned by Merton. As will be seen in the next section, there are many factors which make the percentages based on single questions or statements undependable.

Since the ever present response errors are such that an "ideal" scale can never be realized, one wonders how much of a digression from the ideal would be tolerated by Merton. Perhaps he is under the false impression that measurement ever reaches an ideal of perfection. We hasten to add that there is ample evidence to support Merton's position that one of the chief defects of the Thurstone scales is the scattering of the endorsed statements over a wide range. In Dudycha's 1943 study (32) of the Peterson attitude toward war scale, it was found that those scoring most unfavorable toward war too frequently endorsed statements with scale values at the opposite end of the continuum. Dudycha concluded that the presence of several different continua would explain this scatter of endorsements. He might have added that the spread was probably associated with the none too satisfactory reliability of scales constructed by the equal appearing intervals method. That more than one continuum may be involved in scales constructed by this method is suggested in the report (abstract) by Ferguson (38) who found by factor analysis that more than one component is necessary to explain the item intercorrelations on the war scale. Detail for evaluating this study is not available.

A 1942 paper by Ericksen (33) has been cited by Dudycha as indicating that something is wrong with the Peterson war scale because it did not reflect group differences and changes which seemed logical, and which item analysis seemed to bear out. Ericksen's failure to find statistically significant differences between groups was due to erroneous computation of the standard errors of the differences between means. Instead of using the standard errors of the separate means, he used the standard deviations of the score distributions; hence all of his nine reported critical ratios tend to be quite low, .32 or less. With the use of the proper formula, some of the logically expected differences are statistically significant. Since a part of Ericksen's criticism is based on changes in responses to three supplementary questions as contrasted with the supposedly insignificant change in scale scores, occurring between May 1940 and February 1941, it is of interest to look at the three questions which indicate "a marked increase in the militaristic feeling between the two testing periods." The questions were: "Do you think

the U. S. will enter the present war?" "Do you think the U. S. should aid the allies in every way except by sending troops?" "Should the U. S. fight to protect the Western Hemisphere from a German invasion?" One might grant that a die-hard pacifist would respond negatively to these questions, but does an affirmative response really indicate militarism? Considering the events between May 1940 and February 1941, does a change to more yeses to these questions indicate greater militarism or just good old fashioned realism regarding defense? It may be that Ericksen's general conclusion that the Thurstone type of attitude scale "is too general and abstract in its reference to be practical or valid as a measure of contemporary specific war attitudes" is justified, but not on the basis of his data.

As in all attempts to elicit verbal responses, there are certain rules to be kept in mind when statements are being prepared for possible inclusion in an attitude scale. The interested reader will find the paper of Wang (**125**) quite useful. The tedious sorting work involved in the method of equal appearing intervals can be shortened by a scheme set forth by Seashore and Hevner (**115**) or by a graphic rating method proposed by Ballin and Farnsworth (**2**). That scale values determined by the Thurstone technique may shift has been shown by Farnsworth (**37**) who found significant changes for items originally scaled in 1930 and rescaled in 1940. Even though the rank-order positions of the two sets of values correlated to the extent of .97, the displacement was such as to produce a difference in means for 256 cases scored on the basis of the old and new values. The difference was .69, which yielded a phenomenal critical ratio of 34. Evidently there may be some danger in assuming the same meaning for scores from year to year.

The Thurstone scaling method as used by him and his students has always involved scaling for some particular attitude. Remmers and Silance (**102**) proposed in 1934 that the method could be adapted for the purpose of constructing "generalized attitude scales." Beyle (**4**) had made a similar proposal in 1932. This assumes that a series of statements can be scaled as applicable to a class of objects or phenomena, such as races, or school subjects, or proposed social changes—the "any" type of attitude listed in the first section. Thus one sorting and subsequent scaling provides a set of statements which serve for measuring attitude toward, for example, any institution such as war, marriage, communism, Sunday observance, and divorce. The advantage of the generalized scale is supposed to reside in the fact that many attitudes can be measured without the great expenditure of time otherwise required for sorting and scale construction.

To make the scale applicable to any one of the phenomena in a class, the respondent is simply told that he is to keep in mind the particular attitude object as he checks the statements. That this procedure can become quite ridiculous is illustrated by the following items: "[war or marriage] does not consider individual differences"; "[marriage] gives

too little service"; "[marriage] satisfies only the most stupid with its services"; "[marriage] offers opportunity for individual initiative"; "[war] is too conservative"; "[war] is too changeable in its policies"; "[marriage] appeals to man's lowest nature" (**99**, pp. 18–36).

In order to establish the validity of the generalized attitude scaling technique, one of Remmers' students (**99**, pp. 98–109) constructed a scale specifically for attitude toward teaching and then correlated the scores therefrom with scores based on the generalized attitude toward any occupation checked for teaching. The correlation corrected for attenuation was only .58, but this is said to "show a satisfactory validity" for the generalized scaling scheme even though the validity of the specific scale is unknown. Remmers and his students (**99, 100, 101**) have practically placed attitude scaling on a mass production basis, and have proceeded to use the scales as tools in research on a variety of topics. Individual citation of these studies would add more than 40 titles to a bibliography on attitude research. The fact that the dozens of reliability coefficients reported tend to have a median value of about .70 with values as frequently below .50 as above .80, and ranging down to .07, would indicate that a great deal of uncritical energy has been expended. These reliabilities are lower than those usually found for the Thurstone scales even though from 50% to 100% more items are included in the generalized scales. Evidently the latter scales contain dead timber.

In 1932 Likert (**77**) published a paper in which he presented a method for developing scales for attitude toward internationalism, the Negro, and imperialism. Although this is not a rational approach in the sense that Thurstone used the term rational, the end results are sufficiently satisfactory to justify a sketch of the technique used by Likert. Items or statements which call for checking one of five responses (strongly approve, approve, undecided, disapprove, strongly disapprove; agree-disagree may be substituted for approve-disapprove) were prepared on the basis of a priori judgment. At first, sigma scoring weights for the possible responses were determined by converting the proportion of times each response was checked into the corresponding sigma value on the base line of the unit normal curve. That this laborious conversion job and the resulting cumbersome weights were unnecessary became quite evident when the simple, though arbitrary, scoring weights of 1, 2, 3, 4, 5, for the respective responses were tried out. Scores by this simpler method yielded correlations with scores by sigma scoring which were in excess of .99 and, as one would anticipate from this fact, there was no noticeable difference in reliabilities.

The final selection or elimination of items did not depend upon subjective judgment, but upon the criterion of internal consistency, that is, each item had to show appreciable correlation with the total score on all the items of a scale. The three final scales contained 24, 15, and 12 statements, and the odd-even reliability coefficients ranged from .79 to

.92 while test-retest (30 day interval) reliabilities ranged from .79 to .91. Considering the shortness of the scales relative to the usual 20 to 22 statements in the typical Thurstone type scale, these reliabilities are fairly satisfactory. Eighteen of the 22 statements of the Thurstone-Droba war scale could readily be responded to in terms of the five agree-disagree responses mentioned above. Scoring on the 1, 2, 3, 4, 5 basis increased the reliability to .88 as compared to .78 secured by scoring all 22 items by the Thurstone method.

In further study of the possible advantage of this scoring scheme, Likert, Roslow, and Murphy (78) administered 10 Thurstone scales to several groups. The subjects were instructed to check with a + the statements with which they agreed and to circle the + if they strongly agreed (this much permitted scoring by the usual median check method of Thurstone); the statements with which they disagreed were marked by a −, which was circled if they strongly disagreed; if undecided, they used the question mark. All blanks were then scored by the two methods. The form versus form reliability coefficients were higher for each of the 10 attitudes for all the groups when the scoring was by the Likert method. Actually the reliabilities were appreciably higher—in the .90's as compared with the .80's. Furthermore, the correlations obtained between scores by the two methods were as high or nearly as high as the respective reliabilities would permit. This study would seem to be conclusive evidence in favor of the Likert simple scoring method. Nowhere does it involve scale values (dubious to some) as determined by the method of equal appearing intervals. Does such a scheme constitute measurement? Yes, providing it can be shown that a single continuum is present and providing it is understood that by measuring we mean only rank ordering of individuals.

The foregoing discussion, of course, does not provide grounds for concluding that the Likert technique is superior to Thurstone's scaling as a method for *selecting* items. That the reliabilities for scales built by the Likert technique do not greatly, if at all, exceed those attained by equal appearing interval scales is evident from the reliabilities of .79 to .92 reported by Likert for his own three scales. If two of his scales had been as long as the typical Thurstone scale, the reliabilities would have been higher. However, in the work of Rundquist & Sletto (112) it is found that the reliabilities for their scales of 22 statements (same length as Thurstone scales) are only .80 to .85.

Ferguson (42) has reversed the study of Likert, Roslow, and Murphy by starting with Rundquist-Sletto scales, which were constructed on the basis of the internal consistency criterion used by Likert. For four of these five scales, Ferguson found that the items were not well distributed over a continuum when scaled by the equal appearing intervals method; the items tended to pile up at either the favorable or the unfavorable end of the continuum. He found that the Rundquist-Sletto conservatism scale, the only one with scale values well enough

distributed to permit the Thurstone type of scoring, yielded a correlation of .70 for the two scoring methods. From lack of well distributed scale values and this correlation Ferguson concluded that the Likert technique "does not obviate the need for" scaling by the Thurstone method. But this does not follow from the correlation, in the absence of any information regarding the homogeneity of the group upon which it is based and the reliability of the scores obtained by the median scale scoring; and it can be argued that it is unnecessary to have statements which yield well distributed scale values.

The writer is inclined to believe that some combination of these two competing techniques for scale construction would be better than either one alone. It would seem logical to expect that more reliable scales would result if the Likert method were modified to assure the selection of some items in the middle range of the favorable-unfavorable continuum, or if the equal appearing intervals technique, along with internal consistency, were used for item selection and the median check scoring were dropped in favor of the simpler scoring technique of Likert. Both methods have merits, and both have defects which might be overcome by a combination of the two. As indicated earlier, the Thurstone technique may not always lead to the selection of items which functionally fall along a single continuum, and it can be demonstrated algebraically that the internal consistency criterion alone is not a sufficient condition for selecting items which belong to a single dimension. Using both should help to realize this goal. The adoption of the simpler scoring scheme might allay some of the objections to the "scale" units involved in the Thurstone method. Some might be fearful that the neutral point be lost—this might be good riddance—but it is possible to specify neutral scores when the scoring is by the Likert scheme. A neutral score is earned by a person who tends to give an "undecided" response to statements even though they are piled up at the ends of the favorable-unfavorable continuum, or by a person who is totally inconsistent in his responses. Extreme scores are, of course, obtained on entirely different bases: in one case by the endorsement of one or very few of the statements at one extremity, in the other by strongly agreeing (disagreeing with items of reversed wording) with all or nearly all the statements which support one side of the issue. Yet both techniques yield scores which are distributed over many values, nearly always in a unimodal fashion, both skewed and symmetrical, with perhaps a greater tendency toward symmetry in the case of the Likert technique.

We will next consider briefly other methods, and modifications, which have been proposed for the measurement of attitudes. The attitude of a person may be rated by an individual well acquainted with

him, or as a result of intensive interviewing. The latter will be discussed later in connection with problems of administration or data collection.

Self-rating has been used by Ewing (35) and by Riker (104). The former found rate-rerate (one week interval) reliabilities to be in the vicinity of .80, while the latter found that rate-rerate reliabilities for attitudes toward 13 different issues ranged from .50 to .86, and for intensity the values ranged from .66 to .99. This last figure is suspect. Riker found that, for attitudes toward the Negro, the Germans, the treatment of criminals, capital punishment, evolution, and communism, the correlation between self-ratings on an 11-point graphic scale and equal appearing intervals scales ranged from .55 to .84. If data were available for correcting these for unreliability, there is reason to believe that they would be raised to the high .90's. The correlations between the self-ratings for intensity and the Thurstone scale scores ranged from .57 to .75; corrected, these would likely approach .90, thus proving that the intensity factor is fairly well incorporated within the Thurstone scores. The self-ratings for favorableness and for intensity yielded correlations from .77 to .96, which may be spuriously high due to the possible correlation of errors. It would seem from Riker's results that the self-rating scheme has possibilities for securing a fairly satisfactory distribution of the attitudes of a group when a long scale is impractical. The question whether individuals are thinking of the same continuum, therefore rating themselves on a single dimension, would be difficult to answer.

The selection of items which yield differences between groups with known attitudes has been utilized by Zubin and Gristle (133) in constructing a test for measuring militarism-pacificism. The group taken as militaristic was an ROTC unit at the College of the City of New York, and the pacificists were members of peace organizations. It would seem that such a scheme should automatically lead to the inclusion of items which are valid, hence the validity of a scale so made up should be assured. There is no assurance, however, that items so selected will fall on a single militarism-pacificism dimension; the differentiation for an item might be due to a difference between the groups for a variable which is correlated with militarism-pacificism. A priori judgment in assembling items would help forestall such a happening. These investigators found that the split-half reliability for their 65-item test was .95. In this procedure involving the use of criterion groups, much depends on the real attitude of the individuals belonging to a group—why do they belong? For instance, at Stanford the use of the ROTC unit as a militaristic group might very well lead to the selection of items which aligned individuals along a continuum with pacificism (maybe) at one end and the "liking of horse back riding with the fair sex" at the other (supposedly militaristic) end.

In a study of the attitudes of labor union members toward a central

theme, "the conception of workers as a potentially dominant class," Newcomb (95) discarded the Thurstone method as laborious, and inferior in reliability and in validity when scaled by a group different from the one with which it is to be used (evidently he was here ignoring the futile attempts of others to demonstrate that the scaling group's attitudes affect the obtained scale values). He also rejected the Likert technique as too demanding on the respondent with its consideration of both the statement and the five possible responses. Instead, he prepared 16 pairs of statements, each pair differing with respect to the central theme. The respondent was required to check the preferred item of a pair, and double check if strongly preferred. The writer is unable to see that choosing between two statements with two possibilities for checking preferences is any less demanding on the intellect than responding to statements of the Likert type. Furthermore, no criterion was used for eliminating faulty items, and there is nothing about the procedure which would indicate that a single continuum is involved. That this also results in inadequate reliability is shown by the fact that the median split-half reliability (stepped up) for 13 groups was only .81; for the heterogeneous combined group the reliability was higher, .91. The method used by Stagner (118) in 1936 resembles the Newcomb plan of item selection and elimination in that there was no plan. Thirty-five statements regarding Fascism were assembled as a test, the score being simply the number of pro-Fascist statements endorsed. The split-half reliability was .77 which is certainly below an acceptable standard, especially when the number of items is noted.

A technique which might overcome some of the invalidity of the verbal question-response type of test has been proposed by Proshansky (98). By an adaptation of one of the projection techniques, the Murray Thematic Apperception Test, he secured scores on attitude toward labor for two small groups which correlated .87 and .67 with the Newcomb questionnaire mentioned above. No reliability coefficients were reported, but from these two correlations, based on rather homogeneous groups, it can be inferred that the scheme yields fairly reliable scores. The possibilities of this type of attitude testing should be explored further. The above correlations do not, of course, prove the validity of either the projective or the Newcomb technique.

Rosander (108) has compared two scales, both constructed by the method of equal appearing intervals but differing in that the "behavior scale" contained verbally stated situations with instructions to endorse any of several reactions to each situation, while the "opinion scale" contained the usual type of statement. Both scales were designed to measure attitude toward "social equality of the Negro and the White." The reliability coefficients of the opinion scale were .72 and .90 and of the behavior situation scale, .74 to .87. The two types of scales correlated to the extent of .68, .81, .81, and .89, corrected for attenuation, thus indicating that they measure much the same thing. The author took

these coefficients to indicate validity, but a far better indication of validity is the fact that northern and southern students were rather sharply differentiated by both scales.

Pace (**97**) has also used verbally stated situations calling for a verbal response as to which of several actions the respondent would take. The items were not scaled, but 10 judges ranked the stated actions as to social-political-economic liberalism. Arbitrary scoring weights were assigned to responses. As regards validity, the total scores differentiated completely (no overlap) between 25 known radicals and 25 known conservatives. Odd-even reliability coefficients for different groups varied from .50 to .92. Pace argued that ordinary scales have limitations as to validity, yet he went on to provide "further evidence of the validity" of his test by correlating it with the Rundquist-Sletto Economic Conservatism scale. The corrected coefficient was .72, which is rather high when it is recalled that Pace's test included social and political, as well as economic, conservatism. This type of scale is sufficiently promising to deserve further research.

The verbally stated hypothetical behavior situations with verbal responses can be contrasted with actual behavior in situations created for the purpose of ascertaining attitudes from what a person does instead of by what he says he would do, that is, his verbal responses to verbal statements. May and Hartshorne (**83**) used such a scheme for inferring attitude toward honesty.

A new approach to scaling is found in a 1944 paper by Guttman (**52**) who proposes a rational scheme, based on matrix algebra, for selecting items for scales to measure any type of psychological trait. The chief merit of the method, which is simple in application, is that it leads to elimination of items which are not on the principal continuum, thus assuring that a single dimension is involved in the retained items. Heretofore such assurance could be obtained only by factor analysis methods, which are so laborious as to discourage their use. Neither from the published paper nor through conversations with Dr. Guttman has the writer been able to see that this method adequately takes into consideration the ever present response errors or that ordinary standards of reliability can be attained by so few items as Guttman seems to think will provide an adequate scale. The troublesome validity problem is not solved by this "scalogram" technique, but its superiority on the single dimension problem, plus the safe prediction that reliability can be realized by taking sufficient items, makes this a most promising scaling method.

The reader will have noted that we have not had much to say about the validity of scores obtained by the use of scales. This void is mainly due to the fact that little has been done to establish validity. It is true that many comparisons have yielded differences in attitude scores in the direction expected on the basis of a priori judgment, but anything

like a one-to-one correspondence between verbally expressed attitudes and actual behavior indices of attitude has not been found; in fact, seldom have investigators attempted to ascertain the *degree* of relationship existing between measures of attitude and criteria external to the scales. This is one place where more rigorous research is direly needed. The days of assumed validity should have long passed by.

There is ample evidence that scales can be developed which will yield satisfactory reliability, and it is possible to construct unidimensional scales. The writer sees no way of devising instruments which will yield units which are truly equal. Comparability of units can be attained by the use of standard scores or percentile ranks, perhaps by the method of equal appearing intervals, but such units permit rank ordering rather than true measurement. We would argue that we can secure a reliable and valid ordering of individuals on a single continuum, and that such can prove useful in the scientific study of the attitudes of men.

Very little has been said here concerning the wording of attitude statements. Aside from such obvious rules as those set forth by Wang (125), wording is not a major issue in the scaling problem whereas for the single question method much, as will be seen in the next section, revolves around the wording of questions and the type of response asked for. In constructing scales it is now generally, though not universally, agreed that it is better to provide for multiple responses rather than a dichotomy. The tasks set by scaling and by the single question technique are, of course, not the same. It is one thing to rank order individuals; the apparently easier task of assigning individuals to one of two classes is an entirely different matter.

III. Single Question Opinion Gauging

The primary aim of opinion polling is to categorize individuals or their opinions into two or more classes which are conceived as either qualitatively or quantitatively different. Although this is rather inclusive, it is not our purpose here to discuss the problems which are peculiar to election straw polls. What we have chosen to call the fundamental requirements for an adequate attitude scale can be repeated here as the basic requisites for determining an opinion (or attitude) by the use of a single question. We refer to *reliability* and *validity*, both of which terms are used in the restricted sense, and to *dimensionality*. The application of these concepts in connection with polling techniques poses some problems which differ from those encountered in scaling.

In the case of an attitude scale, the *reliability* is easily calculated by correlating form versus form, or by the split-half method. To get

at the reliability (accuracy) of responses to a single question, it would be necessary to ask the question in some alternate form or to reask the same question at a later time. In either case the time interval should be long enough that the respondents do not give the same responses because they remember their previous ones, but the elapsed time dare not be too long lest real, nonchance, changes take place. These obstacles can be overcome, but phrasing questions in alternate forms which are comparable is not an easy task. It will be noted that reliability of responses to single questions does not involve the notion of how accurately an individual is located on a continuum but how accurately he is classified into one of a few categories. Would he give the same response, hence receive the same classification, if he could be asked the same or a similar question under conditions such that he was amnesic for his original response?

In our earlier discussion of reliability it was pointed out that individual measurement is subject to two types of error, variable or random and constant or biasing. It was stated that biasing errors are of little consequence for scaled tests because when using such tests one is seldom interested in an elusive absolute value. Also in scaling, the use of several questions tends to average out the variable errors—increasing the length of a scale by the addition of items tends to increase its reliability. The situation, however, is entirely different when we look at the reliability of responses to a single question. The variable errors involved in categorizing an individual cannot be readily reduced by increasing the number of questions—the use of a battery runs right into the problem of scaling. It is true, of course, that the variable response errors to a single question will tend to be canceled in calculating the percentage of a group that holds a given opinion. The effect of these errors can be reduced still further by asking the group several questions on the issue, computing the separate percentages, then taking their average. This should yield a more dependable over-all percentage, but it would not provide a more reliable classification for an individual unless his answers were somehow combined, which again involves scaling. The problem of individual accuracy is of considerable consequence in any type of analysis whether it be for ascertaining the correlates of opinion or for the cross classifying of opinions. Obviously, if the obtained responses to a question were entirely chance determined, there would be no hope of finding correlates; and the extent to which chance errors are present limits the degree of correlation. In analyzing interrelationships, there is no use to handicap oneself unnecessarily with response errors, particularly with errors of unknown extension.

There are common sense ways by which single question reliability

can be increased. Other things being equal, a question which is stated simply in easily understood words will tend to yield a reliable reply. Ambiguity, double negatives, and long questions tend to produce unreliable replies. Other things being equal, the greater the respondent's familiarity with the given issue, the more reliable the reply. The greater the personal relevance of the issue, the more dependable the response. The better the crystallization of opinion, the more reliably determined. It cannot be assumed that individual opinions about every issue can be reliably gauged but it is the definite responsibility of opinion pollers to determine *how* reliably individuals can be classified as to their opinions. This aspect of reliability has been practically ignored by the users of the single question technique, although they call their work "scientific."

The only information on the reliability of opinion questions is in Cantril's *Gauging Public Opinion* (18), and this pertains only to one question: "Do you think Roosevelt is doing a good job, only a fair job, or a bad job running the country?" For interview-reinterview, with a three weeks interval, the percentage of identical responses given by 286 persons was 79, which for the given situation tends to approximate a correlation coefficient of .90. This question is said to be one of the most stable, hence it may be inferred that other questions would yield lower reliabilities. By the same method 87% of the same group gave consistent answers to a question concerning whom they voted for in the last presidential election, and information given about car ownership (car or nor car) agreed only 86% of the time. Evidently the factual type of question is none too reliable, so one wonders just what the story is for opinion questions.

A paper by Hayes (56) throws some light on reliability via the consistency of responses to positively and negatively worded statements about the same issues. Regarding opinion concerning two issues, armaments and war debts, the degree of consistency as measured by the tetrachoric correlation coefficient varied from .60 to .70; regarding government ownership, taxes on risks, tariff, unemployment relief, and veterans relief the coefficients ranged from .40 to the .60's; for five other current issues the agreements ranged from the .30's down to .10. Strictly speaking, the use of positive and negative wording may not satisfy the requisite of comparable alternate forms, so the given coefficients may be underestimates of the reliabilities.

When we turn to the constant or biasing type of error, we find that the pollers have an appreciative awareness of the disrupting influence of bias, which is of paramount importance when the aim is to report an absolute—the percentage of a group who favor a certain action or who believe in something. To do this, it is necessary to avoid biasing errors, a subject to which we shall presently return.

The second basic requirement, that of *validity*, does not pose any problems unique to the single question technique. Does the question elicit a response which reflects the individual's opinion on the given issue, or is it a reflection of his opinion on some other issue? Is there any evidence that the verbally expressed opinion corresponds to a behavioral indication of opinion? Can a man's response be regarded as truthful? All this boils down to whether a person's opinion or attitude can be safely inferred from what he says. Or, to state the problem differently, it can be said that validity is a matter of interpretation. Suppose a soldier says he would like to go overseas soon, and that this is interpreted as indicating a desire to get into active combat. At once the question of validity faces us. Does his statement really indicate high fighting morale, or does it merely indicate a desire for change or for adventure? His verbal expression could be reliable in that he might consistently say the same thing, but to assume that it is a valid indication of a desire to fight may be far from the truth. Opinion pollers have done little to establish the validity of the questions they use. Here is a wide-open field for research, the difficulties of which should prove challenging to the best minds.

The few studies on validity have been concerned mainly with validity in a broad sense, e.g., polling results are checked against voting behavior for a group as a whole. That is, the check does not involve establishing the extent to which individuals voted in accordance with their responses to the poller. Agreement of over-all percentages is meat for the journalistic mill and perhaps sufficient for the crude comparison of groups, but it is difficult to see how a science can be built around opinions without first establishing individual validity.

Blankenship (**7**, **9**) has checked five types of questions, about horse racing and parimutuel betting, against voting intention, and found that the percentage of agreement ranged from 85 to 99. This is said to indicate validity but it is nothing more than agreement between two verbal expressions. One might have expected that Cantril's volume (**18**) would have tackled the problem of opinion question validity but such is not the case. Evidently "the serious problems encountered in every phase of the polling operation" (**18**, p. viii) do not include what the writer considers to be foremost in importance, namely, validity and reliability for the individual response. The volume does touch on some of the factors which contribute to unreliability and invalidity, but a frontal attack is sorely needed. In a 1945 article, Connelly (**24**) has characterized validity as the "real problem" facing opinion pollers. It is heartening that at least one poller has so expressed an awareness of this requisite.

That falsification in answering questions is indeed a fact is evident from the recent note by Hyman (**61**), who found that 17% of 243 indi-

viduals who were known to have redeemed war bonds answered that they had not cashed in any of their bonds. He also reports that the agreement between workers' statements and plant records on absenteeism ran from .88 down to .30, with a median tetrachoric correlation coefficient of .60 for 18 industrial plants.

Our third basic requirement for an adequate scale, *that it shall function in a uni-dimensional fashion*, must be stated somewhat differently as a requisite for a satisfactory single question. The categories into which individuals are classified on the basis of their responses to a single question can be either qualitatively or quantitatively different. The hope is that those who are placed in a class or subgroup are more alike in their opinions (or attitudes) than the total group. To argue that the classes represent quantitative differences is to assume differences in degree on, say, a favorable-unfavorable continuum. This assumption is doubtless true for a very large number of opinions, but in order to make a meaningful inference to the effect that those in one category are more favorable in their opinions, it is necessary to establish as fact that the responses given to the question do involve a single dimension. Otherwise, one might be dealing with an unknown mixture of quantitative and qualitative differences. If the purpose is to categorize individuals according to qualitative differences in opinion, then the greater the qualitative homogeneity of the opinions of those in a class, the more meaningful the classification. Thus one requirement for a good opinion question is that it shall lead to responses which either indicate differences in degree on a single continuum or permit classifying into qualitatively similar subgroups. This means that the question must be so simple that it will be interpreted in the same way by all respondents and that, in so far as possible, it should be answered from the same point of reference. That individuals may give similar responses but for different reasons is an obvious possibility. Intensive research is needed to determine the various reasons back of responses.

It is not here claimed that all the problems which beset opinion polling can be subsumed under the three basic requirements but these have implications bearing on practically everything that has been done and said about opinion gauging by means of single questions. A great deal of research has been directed at the *factors which tend to bias responses*. In fact so much is known about the variations which can be produced, and so little is known about which variation is most nearly correct, that one is apt to become pessimistic concerning the possibility of single question polling ever contributing scientifically useful data.

It is not our purpose to review or criticise the specific studies which

have to do with the general dependability of results secured by the single question method. Rather it is our hope that a sketch of known and possible factors which affect the reliability, validity, and scientific worth of data based on responses to questions will aid in an evaluation of the entire polling procedure.

Of the several variations in ways of ascertaining opinion, the simplest is the direct question which calls for a "yes" or "no" answer. These responses are easy to code and tabulate, and their meaning is encumbered only by the meaning of the question. The assumption that a question has been interpreted in the same way by all respondents is made, usually without being stated explicitly, and is then forgotten, although it is amenable to at least partial check. The factors which may be operative in producing "yes" responses which are not of the same significance are many. The words in the question may carry *different connotations* for individuals of different cultural backgrounds or intellectual levels. For example, a surprising number of Negroes, interviewed orally, seemed opposed to government control of profits. Probing brought out the belief that God alone should exercise control of prophets! If a cross section of the adult population is to be polled, the vocabulary limitations of the subaverage must be kept in mind when questions are being phrased.

Aside from word meaning, the answer to even a simple question is partly dependent upon the respondent's *frame of reference* which may often differ from that hypothesized by the poller and from that of others who give the same response. The importance of the frame of reference is well illustrated by an unpublished check, by The Department of Agriculture's Program Surveys, on a recent Gallup poll. Gallup had reported that the majority of the people want things to remain the same, as judged by their responses to "After the war, would you like to see many changes or reforms made in the United States or would you rather have the country remain pretty much the way it was before the war?" Program Surveys asked the identical question, then probed further to learn how the respondents interpreted "changes or reforms." Sixty per cent had responded in terms of domestic (as intended by Gallup) changes and reforms, for 7% the frame of reference was not ascertainable, and the remaining 33% interpreted the question as referring to one of the following: technological improvements, changes in traditional political-economic structure, foreign affairs, immediate war conditions, personalized conditions, and desirable state of affairs. Of those who answered from the domestic point of view 60%, as compared with 46% of the total sample, voted for "changes and reforms."

Another frequently uncontrolled factor, which is closely related to frame of reference, pertains to the *understandability of the question and/or the issues involved*. For instance, how dependable an answer can one expect to the following question used by the Office of Public Opinion Research in a recent study (12): "For handling domestic problems like unemployment, the converting of war plants to peacetime use, or the demobilization of soldiers—do you think the Government should set up a central agency *now* with full authority to make plans and with full authority to carry out these plans as soon as the war is over?" The writer of that question certainly has an ivory tower notion regarding the intellectual span of the average adult. Questions are also being asked about the postwar world, the future peace, and other rather abstract issues, and the answers are being treated as though coming from a public with opinions well crystallized by experience and information. Such indeed become illustrations of "a consensus of worthless opinion is a worthless consensus of opinion."

Even if a simple yes-no question is clear, calls for the same frame of reference, and has a cognitive level low enough for all respondents, it entails other difficulties which can be subsumed under the *mechanics of questioning*. Some of the factors which may affect the answers are: positive versus negative statement of the question; attempted balancing of positive statement by ending with "or not"; loading by introducing emotionally charged words or phrases; the presence of contingent or conditional ideas; the influence of juxtaposed questions; suggestive elements; alternate wording; prestige elements; personalization of the question; stereotypes; technical words; biased wording; etc. Research directed at these factors indicates definitely that the percentages of yes (or no) responses can be varied. For example, Cantril (18) shows that "interventionist sentiment" between May and September 1941 showed an apparent variation from 78% down to 8% mainly as a function of the questions asked. Which is the correct percentage? No one knows.

Whether systematic investigation of such factors will ever lead to principles sufficiently general to permit their application to new situations is debatable when judged by the rather large amount of research already reported. Further research would serve a useful purpose if it led to more critical interpretation of over-all percentages, but such research cannot determine what particular combination of factors yields the "truth" unless external criteria are first established. Here we reach an impasse—seldom, except in the case of voting behavior, has an adequate criterion been found. It has been argued by some who are well aware of the dangers of taking percentages as absolutes that in comparing the relative standing of groups on an issue one need not worry about

the influence of these factors. This assumes, of course, that any loading leads to the same amount of bias in the groups being compared.

Many objections have been raised against using questions which restrict responses to yes or no. Aside from the fact that the obtaining of information on certain issues may not always be amenable to yes-no questioning, it must be admitted that *qualitative and quantitative differences in opinion may exist among those who give a yes (or no) response.* In particular, there may be marked affective or intensity differences among those who give an affirmative answer; ditto, a negative response. In order to get at possible qualitative or quantitative divergencies among the yes (or no) responses, a number of response categories may be provided. Desirable as this may seem either from the respondent's viewpoint or as a scheme for securing added information, the inclusion of multiple response categories introduces a new set of methodological problems.

Possible variations in the response setup include the following: two, three, four, or more categories; extreme categories (effect of superlatives); omission or inclusion of a definitely labeled neutral response and varying forms therefor; unbalanced toward positive versus toward negative versus balanced categories; inserting an "undecided" response; reversing order of categories; including additional words in the possible responses; lengthy statements as responses; etc. Research has been directed toward discovering the effects of these and other factors at the response level. As in the case of question form, the results indicate that variations in the response structure lead to variations in classification and interpretation, but there would seem to be no absolute basis for determining the optimal response setup.

Other variations are being used. Sometimes the interviewee is asked to check an item or items in a list or to rank them in importance. The order of the items in the list may be a disturbing factor here. In some cases the respondent is asked to express his preference for a number of paired items. This, too, has its difficulties unless it is feasible to use a complete paired comparison scheme. At times, the so-called open-end or nondirective type of question is used, thus permitting a free or unrestricted response.

The open-end question coupled with highly trained interviewers has been advocated, particularly by Likert of Agriculture's Program Surveys, as one way of overcoming some of the difficulties arising from variation in question and answer form. The question is stated in the simplest possible manner and followed by further questioning to determine whether the respondent really understands the question and the issue, and to ascertain whether he maintains his opinion under mild cross

examination. In other words, it is the job of the interviewer to be sure of the respondent's comprehension and to help him formulate, but not form, his opinion. Then on the basis of the interview, the respondent's opinion is categorized either by the interviewer or by an analyst or coder who studies the verbatim record.

The nondirective question, as frequently used in self-administered schedules, has the advantage that answers are not suggested by the listing of possible responses, and the disadvantage that the elicited response may represent what happened to be recalled or thought of at the moment. Difficulties are also encountered in coding or classifying the free responses. This type of question is most effective in the interview situation by virtue of the possibility of probing further by additional questions. Since these questions must be tailor made on the spot to fit the peculiar requirements of a given interview, it follows that such an interview cannot be standardized. Much is left to the ingenuity of the interviewer, and a high degree of training or an inborn knack becomes a prerequisite in order to objectify the interview situation. Interviewer bias is apt to be an uncontrolled factor.

If that bias can be overcome by training and proper supervision, and if the coding from the verbatim record of the interview can be objectified, it seems reasonable to believe that in general the use of nondirective questions will lead to more dependable results than those obtainable by restricted responses, especially when opinions on relatively complex issues are being sought. It is generally agreed that the open-end approach is of particular value during the preliminary phases of most studies, but there is a lively controversy, not entirely free of emotional cultism, regarding the proper place of free response interviewing as a method of ascertaining opinion. The extent to which the free and restricted response methods yield similar results is in need of further investigation. The supposed greater dependability of the open-end method must be weighed not only against certain subjective factors already mentioned but also against its much greater cost per case. It is more time consuming than ordinary interviewing, which in turn is more costly than self-administered schedules (not often feasible), and the work of the coder is slower. The coder also needs more training to overcome another source of bias and possible misjudgment.

The process is so slow that for certain governmental and journalistic purposes the reasons for making a given study may have vanished before the report is available, and so expensive that it is difficult to secure sufficient cases to permit breakdowns necessary for the analysis of relationships. Without doubt, however, the utilization of open-end questions by well trained and intelligent interviewers could serve as a

useful check on responses obtained by more formal interviewing or by self-administered questionnaires. One would expect superiority for the open-end method when dealing with issues which for various reasons are not clear-cut, but for simpler issues this technique may not yield results any more dependable than those obtained by typical polling or questionnaire procedures. The devotees to interviewing by nondirective questions argue that this method can and should be used to determine the best combination of question form and response setup. (For a more detailed discussion of the open-end method, see references **73, 80, 117.**)

It takes no stretch of the imagination to see that by taking all the ways in which a question may be stated and all the possible variations in the response setup, it is possible by permutations and combinations to outline an enormous number of so-called research projects with the aim of studying the effects of purely mechanical aspects of asking questions and securing responses. A perusal of the literature and conversations with opinion researchers in government indicate that such mechanical research is regarded as of first importance in methodological studies. This may be due to the relative ease with which such research can be done and the fact that the investigator is fairly sure of obtaining positive results, in the sense of statistically significant differences. In the absence of any absolute or external criteria for judging which of many variations is best, and in view of the fact that opinion pollers are continuing to report percentages as though such figures had absolute meaning, it is difficult to see the value of this mechanical type of methodological study. Semantics and related disciplines may profit therefrom. Enough research along this line has already been done to indicate the fallibility of single question polling, and to provide caution signs for careful pollers, so it is doubtful whether more of this type of research is worth while.

Now that it is known that over-all percentages do vary with different ways of stating a question and with alternate response setups, what can be done about it? Obviously a percentage based on one form becomes a crude approximation. To know that another form will produce a lower or higher percentage does not constitute a safe basis for deciding the direction and extent of the possible bias produced by the particular form which has been used. In the absence of information concerning which of various alternatives involves the least bias, it might be assumed, perhaps falsely, that the least biased is that which gives a percentage near the average yielded by several forms. Preliminary work would be needed on each issue in order to learn which form was best by this criterion, and since the total possible number of forms is large, this task is scarcely feasible.

Another method which should yield about the same over-all percentage as obtained by this "average" question-answer form, and which rests on a similar untested assumption, is the split-ballot technique. This consists of preparing the schedule in several alternative forms and administering them so that the first interviewee answers on the first form, the second on the second, and so on through the sixth if there are six forms, then the seventh on the first, etc. For each question-answer form, some response is regarded as indicating, say, a favorable attitude; hence a count of the individuals, regardless of the form used, allows the computation of a percentage for the entire group. This over-all percentage will correspond exactly to the average percentage for the six forms if the number of cases (N) is the same for each form. Although a percentage arrived at by the split-ballot technique may be biased, it seems only natural to have more confidence therein than in the result from a single form. The public opinion pollers have, without doubt, ironed out some bias by this technique and thereby reported more dependable percentages.

That such a procedure does not end all the troubles which beset the pollers should be obvious. As already indicated, this type of average percentage may also be biased, and the inclusion of all possible forms of question and response setup, even if feasible, would not guarantee the elimination of bias. If breakdowns on such variables as age, sex, socio-economic status, rural-urban residence, party affiliation, etc. are to be made, it is necessary to distribute the different forms evenly within each subgroup. This would not be impossible for a large polling agency, although if many forms were necessary, and such would seem to be the case, the difficulties would be great. All this pertains to reporting either one over-all figure, or percentages for various subgroups. Such may meet the requirements for reasonably dependable reporting for the press, for use by governmental agencies, and for the routine comparison of groups; but a science of public opinion will not make much progress without investigating the interrelationships and antecedents of opinions (or attitudes). Do those who have a given attitude on one issue tend to express certain opinions on other issues? What are the common psychological and sociological characteristics of those who are classified as having the same opinion on an issue? How do they differ from those who hold an opposite attitude?

To answer such questions it is necessary to make the breakdowns on the basis of the expressed opinions, but for this purpose the split-ballot technique is obviously inappropriate. In fact, its use for an issue upon which a breakdown is to be made would guarantee that those classified as, say, favorable on the issue would really be heterogeneous

in their opinions thereon. For such analyses it is necessary to stick to one form of the question-answer setup, and naturally the one form to use is that which is least biasing, if this is known. And so we come right back to the fundamental problem of ascertaining how to select that question-answer form which avoids biasing the results. This, we repeat, cannot be done by demonstrating what seems obvious—that percentages can be made to vary by varying the form of the question and the permissible responses. An external criterion is absolutely necessary, but even if such were available one would be faced with additional complications. Suppose that various question-answer setups are to be tested against a known percentage, and it is found that a particular setup yields this percentage exactly. From this it would be concluded that a nonbiasing form has been found. What is the generality of such a conclusion? None. One wouldn't know whether this form would be best for another issue, nor whether agreement with the known percentage meant the absence of biasing for subgroups resulting from breakdowns by age, education, etc. The latter could presumably be checked in our hypothetical case. It would seem that the quest for an unbiased question-answer setup would be interminable.

Another problem which seems to perplex some critics of the polling techniques has to do with *intensity*. It is argued that it is not enough to know how many people favor a given proposition; we need to know something about the intensity of their attitude. Do all those who favor a thing do so with the same degree of affect? Are proponents or opponents of a social question more intense? The degree of emotion involved in the holding of an attitude or the expression of an opinion is of twofold significance: its bearing on action if action is possible, and on the stability of opinions. Presumedly, the greater the intensity the greater the likelihood of action, and the less the intensity the greater the possibility of changes in opinion.

Several schemes have been used to get at the intensity factor. Cantril and Rugg (**21**) have asked respondents to indicate whether their feelings were strong or mild, and Roper (**106**) indicates that he taps intensity by a scale of possible responses (strongly favor, favor, etc.). The most serious attempts to measure intensity, the studies reported by Cantril (**18**, pp. 51–65), will be reviewed briefly here.

One study involved two ratings of intensity, by self and interviewer, into one of five classes: strong approval, mild approval, no opinion, mild disapproval, strong disapproval, of Roosevelt's message to the dictators. It was found that both ratings were related to the respondents' opinions of Roosevelt as shown by their answers to other questions. The data do not permit a statement as to how high the relationships

were. Next a graphic thermometer device for intensity self-rating was studied with reference to voting intention. The "very strong" supporters of a candidate differed from the "weak" supporters in their endorsement of statements regarding the candidate. Then a graded scale of only three statements, constructed by the Thurstone method, was utilized. The differentiability of this device proved to be so inadequate that the present reviewer completely rejects the statement that the "experiment demonstrates that simplified forms of intensity devices can be effectively used on public opinion polls" (18, p. 59).

Finally a far more extensive study is reported. This time seven devices were used, four of them under controlled conditions so as to determine their relative discriminating power. These four were a verbal self-rating of strength of feeling, self-rating on the graphic thermometer scale, a four-step logical or attitudinal scale, and verbal self-rating of certainty or sureness. The predictive efficiency of the four devices was to be determined by the extent to which intensity so measured on four key issues would predict the respondent's answer to certain test questions. Two of the key issues failed to function as expected, but the failure is rationalized, after the fact. For the two remaining key issues, percentage tables are presented for a total of five test questions, and from these tables each device is characterized as very good, good, fair, or poor. Since the criterion for this is not stated, and since one criterion which the present writer has applied to the tables fails to yield corroboration, the conclusions in Cantril's book are suspect.

Incidentally, this study is a good example of inadequate statistical treatment, and failure to report data in such form that any one can remedy the treatment. True, percentage tables are given (18, pp. 306–310) but the needed N's are missing, so our evaluation of the relative predictability of the four devices may be inadequate. At any rate, when one uses a simple criterion, the absolute differences between the percentages for agree versus disagree (or yes versus no) on the test questions, one finds, for example, that when the key question concerned trust in Russia, the two devices characterized as "very good" do not rank as high as the other two, one of which is said to be "good," the other "fair." For the key question on nationalizing industry our simple index places two devices labeled "very good" below a device labeled "good." We make no claims for such a simple index, but the reader of Cantril's book has a right to know the basis for the ratings given the four devices. To characterize predictions as very good without presenting a measure of association so that one can tell how good "very good" is injects an avoidable element of subjectivity into scientific research.

In this section we have attempted to point out some of the methodo-

logical difficulties associated with the study of opinion or attitude by way of the single question. We have found that investigators have practically ignored the problem of the reliability with which an individual's opinion is categorized. Nothing is known concerning the variable errors involved in opinion gauging. The question of error due to bias in question or answer form as affecting over-all percentages rather than individual responses has been studied by Blankenship (**6, 7, 9**), Cantril (**15**), Ghiselli (**50**), Hayes (**56**), Link (**79**), Matthews (**82**), Roper (**106**), Roslow, *et al.* (**109**), Rugg (**110**), Rugg and Cantril (**111**), and others. These studies merely demonstrate the important fact that variations in percentages can be produced by varying the question and/or answer form, but which setup is least subject to bias has not been ascertained. This ignorance concerning unreliability due to bias and to ordinary variable response errors, plus an equal ignorance of validity in the restricted sense, and scant knowledge as to qualitative or quantitative homogeneity of those assigned to an opinion category, to sav nothing of the problem of being sure that issues are understood and opinions crystallized,—all these constitute a serious indictment of the claim that a science of public opinion is being or can be built around the single question procedure. The use of interview schedules with interlocking questions will not strengthen the edifice unless the responses to the questions are somehow combined into an index so as to obtain a more dependable basis for classifying individuals. Again the problem of scaling confronts us.

Not all of the difficulties in public opinion polling have been discussed here. Problems of interviewing or schedule administration, of sampling, and of analysis, have not yet been considered. Before going on to these, we shall make some suggestions as to what might be done about the grave inadequacies of single question polling.

I. *Reliability* (individual consistency or inaccuracy of classification due to variable errors) *can be readily studied, and therefore should be*, in order to learn how high, or low, it is and also which question-answer form is least subject to variable response errors. It is the writer's belief that the reliability for single questions will be found to be low; if measured by some coefficient equivalent to product moment correlation, we estimate that typical reliabilities will be in the .50's or .60's, sometimes lower and occasionally higher. Furthermore, it is unlikely that any amount of careful work will lead to an appreciable improvement in reliability over that attainable by using the question-response setup which turns out to be the most reliable.

II. Until rigorous research is done on validity, the pollers and the social scientists who use the single question method can expect continuing criticism regarding whether the questions used get at the opinion

intended. *Validity studies can and should be undertaken.* Research on reliability and validity is, of course, humdrum work which is not productive of spectacular titles for articles and books, but some of the problems associated with determining reliability and validity should prove challenging to the most ingenious minds.

III. When one turns to the question of *whether one can be sure that responses are qualitatively or quantitatively comparable*, one is faced with a problem the solution of which has not, to the writer's knowledge, been found.

IV. The most difficult problem of all, however, has to do with *finding the form of question and form of response setup which is least subject to bias.* No more research is needed to show that the question-response setup is a source of possible error, the magnitude and direction of which cannot be established unless external criteria are available. This is seldom the case, so the bias in over-all percentages cannot often be determined. We have already indicated that polling has reached an impasse on this, and that the chase after absolute percentages is a delusion and a snare.

As the writer sees it, there are at least two ways in which social scientists can avoid this cul-de-sac. Both ways are relatively expensive, but fewer and better studies are perhaps desirable. The researcher might be forced to get along with fewer cases, and therefore be unable to make as many breakdowns—perhaps that would be a blessing too in that he would be forced to avoid some of the currently and routinely made breakdowns which seem to contribute so little to the understanding of opinions and attitudes.

The first way out is suggested with some trepidation as it has been the center of heated, though unpublished, controversy, which might have been avoided had the personalities involved been less uncompromising. This suggestion is *that greater use be made of the open-end, nondirective, intensive interview technique.* We have already discussed its merits and demerits. It is generally agreed that this method is of particular value during the preliminary or pretesting phase of most studies. It is the writer's belief, subject to correction by research, that the results obtained by highly trained interviewers using nondirective questions will possess greater reliability (accuracy for the individual), greater validity, greater homogeneity for those assigned to an opinion category, and less bias than results obtained by the ordinary polling procedure. If it can be shown that for a given issue the ordinary method yields data just as dependable, judged either by external criteria or by high correlation (say above .95) between the results obtained by the two methods, then the simpler procedure should be used. As to the intensity factor, one might expect the interviewer to be able to make an

adequate appraisal thereof. It is not claimed here that the general dependability of the open-end method has been established—it just seems logical that it should be better than the regular procedure. Obviously, research is needed to delineate further the respective merits of the two approaches.

We have no hesitancy in making our second proposal, namely, *that single question opinion gauging be discarded in favor of opinion measurement by attitude scales.* By using scales a satisfactory degree of reliability (accuracy) can readily be attained. By factorial analysis or preferably by the Guttman (**52**) scaling method, uni-dimensionality can be guaranteed. There is evidence (**104**) that scaling automatically takes care of the intensity factor. Validity would have to be established but this is no more difficult for scales than for single questions, and with higher reliability the validity could be higher. Furthermore, validity is more meaningful when a single dimension is involved. The whole worrisome problem of bias due to question and answer form nearly vanishes. Scaling will be of no particular advantage to pollers and others who are satisfied with reporting percentages as absolutes. In fact, the placing of individuals along a continuum handicaps the percentage addicts. To use distributed scores as a basis for specifying what percentage of the population favor an issue is analogous to using a distribution of heights to say what proportion of people are tall. Scaling assumes quantitative differences—if qualitative differences are to be determined, scaling would be inapplicable.

There are additional advantages to measuring opinions by scales. By their use the social scientist is less apt to be misled or to mislead others with percentages. He has a more accurate instrument for classifying people into categories. He can, for example, segregate the upper, or lower, 30% with some assurance that the included individuals have been fairly reliably placed therein; by breakdowns on other variables he can make as meaningful analyses as are possible with the single yes-no question; he can use the distributed scores in his analytical work and thereby be able to express the closeness of relationships in an exact manner; and predictions can be more easily made from scaled opinions. If the aim is to analyze the variation in opinion on an issue, graduated scale scores greatly facilitate this; the variance can be broken down into component parts, a rather difficult if not impossible task when nothing but percentages are available. Furthermore, if one is interested in comparing groups obtained by different sorts of breakdowns, the analysis of variance technique can be used as the significance test (the requisite assumption of normality of distribution can be met by using transformed scores when necessary). Another advantage of scaled opin-

ions is the fact that the variation within groups indicates the relative homogeneity of groups in their opinion about an issue.

If it is objected that a given problem does not justify the extra work entailed in scale construction and the added schedule administration time, there is reason to suspect that the project isn't of much value from a rigorous scientific viewpoint.

IV. Administration

We shall here consider briefly some of the problems associated with schedule administration, or the collection of attitude and opinion data. Certain aspects of administration are closely related to sampling, which will be discussed later. For the most part, data on opinions are secured by mail, by group testing, and by interviewing.

Mailed questionnaires have two major drawbacks, both of which have a direct bearing on sampling. First, it is difficult to build up mailing lists which are representative. The 1936 *Literary Digest* debacle will be recalled. Secondly, those who return questionnaires are apt to be atypical. Studies by Cahalan and Meier (**14**), Reuss (**103**), Shuttleworth (**116**), Stanton (**119**), and Suchman (**121**) have demonstrated that all too frequently the selective factors involved in mailed ballots and their return are apt to be highly significant. Mailed schedules or the telephone may satisfy the less scrupulous but such methods should be taboo.

The administration of schedules to groups, with each person checking or writing his own responses, is less expensive than interviewing. This is feasible for certain types of study, particularly when one is studying changes in attitudes or opinions produced by experimental procedures *and* when the necessity for securing a representative cross section of some defined population is not so important. For long schedules, self-filling of blanks by individuals in groups is less awkward than is individual interviewing. The group method has been widely used in studies based on college students; the sophomore in elementary psychology or sociology is easily accessible and usually willing to dawdle time on such schedules as are passed out to him. Since the college student is usually able to understand simple questions, there is no reason to suspect that he needs help in marking a questionnaire. As compared with interviewing, self-administration in groups rules out the possibility of interviewer biasing.

Even if it were possible to assemble groups for studying a cross section of the voting population, the write-in method would not always be advisable. Those at the lower educational levels might find the questions too difficult to answer without help. And a query can be raised regarding the comparability of results obtained by write-in and by personal interview. Some unpublished studies of the Research Branch of

the Army's Special Services Division are of interest on this point. Near the beginning of the Branch's work the task of using a long questionnaire as a basis for interviewing a large number of troops looked impossible. Accordingly an experiment was set up for checking group versus interview secured data. Details cannot be given here, but there was evidence on the basis of percentages and scale scores that the two methods yielded comparable results except for those at the lowest educational levels. In the Army it is, of course, fairly easy to assemble groups and at the same time secure a good cross section.

For the great bulk of research in the field of public opinion, the chief method of securing opinion data is by face-to-face interviewing, and this seems likely to continue because of the near impossibility of bringing randomly selected individuals together. Even in the Army situation, interviewing was essential in the preliminary or pretesting stage of various studies. Quite frequently in both Army and civilian survey work it is necessary to begin with tentative questions on the issues to be studied, try them out by intensive interviewing of a relatively small number, then revise, eliminate, or add questions so as to secure those which seem adequate. Perhaps this process must be repeated several times before one hits upon questions which seem satisfactory. Sometimes the poller may be set the task of studying opinion about an issue which is none too well defined or about which so little is known regarding the reactions or thinking of people that it becomes necessary to use the open-end nondirective interview as a means of setting the questions. It goes without saying that this type of preliminary interviewing should be in the hands of trained and experienced individuals, and that much is gained by conferences between interviewers at this stage. No doubt some of the critics of the polling technique would be less skeptical about the dependability of results if they were required to participate in a few rounds of intensive pretesting. At least their subjective outlook on the whole procedure would be different.

The principal problems associated with interviewing have to do with the requisite training and with the possible influence of the interviewer on the respondent. This influence may result in bias or merely a reluctance of the respondent to express himself to a stranger. Those interested in a discussion of the sources of errors in interviewing as a method of collecting factual information will find the little book by T. E. Neely (94) worth while.

The type of individual used as an interviewer and the amount of training and supervision provided seem to vary from agency to agency. Apparently, Gallup depends mostly on part-time interviewers with a minimum of training and supervision. The writer has the impression that the National Opinion Research Center (Denver) and Roper's *Fortune* Poll pay more attention to selection, training, and supervision. Agriculture's Program Surveys, under Likert, employs full-time interviewers who undergo relatively more training and are rather carefully

supervised. Little is known concerning the value of training and supervision. A reasonable hypothesis would be that more dependable results would come from those who are trained and supervised. The chapter on this subject in Cantril's book (18) presents so little evidence as to the difference between the training of the "trained" and "untrained" interviewers that one wonders why the study was ever made.

The extent to which respondents may tend to be influenced by having to give responses to a stranger has been studied by asking a random half of a sample to mark ballots secretly and drop them into a sealed or padlocked ballot box while the other half give responses to the interviewer on identical questions. There is some evidence (3) that more people give an undecided response, on candidates for election, in the interview situation than by secret ballot. The study by Field and Connelly (47) compared results secured by interviewing with those obtained in actual polling stations on election day. Again the secret ballot polled fewer undecided responses, this time on political issues. The responses secured by interviewing checked reasonably well with the "voting" results on two of the three issues, but on the third the difference in percentage of affirmative answers was too large to be chance. Cantril (18) found statistically significant differences on three questions, suggestive differences on four, and no differences on three questions for secret ballot versus interviewed groups. The differences seem to occur on issues on which a particular response carries social approval or prestige. It will be recalled that the experiment cited above, on self-administered schedules versus interviewing in the Army, did not produce response differences. This did not involve the secret ballot method, but schedules were unsigned and interviewee anonymity was guaranteed. Furthermore, it has been the subjective feeling of both civilian and soldier interviewers in the Army that the face-to-face interview situation does not inhibit the respondent. This may not hold for other groups.

A difference in status between the interviewer and interviewee is another source of error. Responses of Negroes to Negro and to white interviewers may differ (18) and the socioeconomic status of the interviewer has possible effects (18, 65).

The researches of Blankenship (8), Cantril (18), Katz (65), and Udow (124) seem to indicate that on certain issues the opinion of the interviewer tends to bias responses. Cantril concludes:

> Although interviewer bias exists, by and large the biases in one direction cancel those in the opposite direction, so that the over-all percentage of opinion is not likely to be significantly wrong. If an investigator wishes to minimize interviewer bias, he should choose an equal number of interviewers who are biased in different directions (18, p. 118).

So choosing the interviewers will scarcely be feasible for even a large agency when the schedule contains questions on more than a few issues. But suppose we stick to just one issue upon which opinion is fairly

evenly divided, will the bias be averaged out by choosing an equal number of interviewers on each side? This depends upon the tenability of an assumption, namely, that the biasing effect of the pro-interviewers is the same in amount as that of the anti-interviewers. That such may be far from true is suggested by the fact that for many issues the supporters of one side are more apt to be evangelistic crusaders.

We are forced to conclude that the effect of the interviewer, whether unrelated to his own opinion (as in the secret ballot studies) or related to his own opinion (biasing), constitutes another source of error in overall percentages. Also of importance is the fact that in the study of interrelationships we have another attenuating error. The possibility of overcoming this source of error by more carefully selecting and training interviewers should be investigated.

V. Statistical Issues

Research on attitudes and opinions necessitates the use of certain statistical techniques as bases for making inferences. Usually the statistical treatment is relatively simple. The typical single question opinion study involves little more than the comparison of percentages for groups, while in the typical attitude study the means for groups are compared and correlations are determined. The statistical requirements are not much different from those in other areas of social science research. In general the value of a given study will depend in part upon adequate sampling, upon freedom from erroneous statistical manipulations, and upon the extent to which the statistical analysis and report are satisfactory. A thorough discussion of these statistical issues is neither feasible nor desirable here, but a brief sketch will serve a useful purpose if it leads only to an awareness of the statistical aspects of research in this area.

Sampling. All attitude and opinion studies are made on samples drawn from a universe or universes about which inferences are made on the basis of the statistical measures which describe the samples. In making such inferences, two types of error must be kept in mind—the random or chance sampling error and the biasing error due to possible selective factors. Actually, one finds here the usual sampling issues: the adequacy of the sampling techniques for yielding a representative cross section of a carefully defined population which itself is of sufficient generality to permit generalization of scientific value, and also the adequacy of sample size for reducing random errors to a desirable minimum for the total sample or for the subgroups involved in the analysis.

First, let us consider the sampling problems associated with the typical study in which attitudes are determined by the use of scales. Such studies are practically always made on college students. The real

issue here is not only that of sampling adequacy but also of universe adequacy. Just what population is being sampled? The population of college students? All colleges? Just one college? Is this one institution typical or atypical? How is the sample drawn from among the students of the single institution? So as to get a cross section? Those who have even casually glanced at the literature will be able to answer these questions. Those who make the studies could also answer, but not without embarrassment. The fact is that all too frequently the investigator turns to the group that is most accessible. This is apt to be his own or a colleague's class of sophomores in either elementary psychology or sociology. If he is interested in comparing freshmen, sophomores, juniors, and seniors he can usually find an instructor who will willingly admit that the progress of science is more important than his Monday lecture. Little time is wasted in defining the universe and drawing a real sample therefrom, but one cannot infer from this that the investigator never heard of sampling. On the contrary, he demonstrates his competence by calculating sampling errors and talking about statistical significance.

It is, of course, obvious that this procedure does not involve the essential steps in sampling. The universe is seldom defined and the subjects are not often drawn in such a way as to constitute either a random or a stratified sample. The formulas used for computing the sampling errors are nearly always those based upon the assumption of random sampling, yet the two most elementary and fundamental principles of random sampling are not followed. That is, the sampling technique is seldom such that each individual in the universe has an equal chance of being drawn and that the drawings are independent of each other. So little is known about possible selective factors that it is even impossible to use the sample to define the undefined universe.

This inadequacy of sampling plus the fact that the groups used are from highly restricted universes leads one to question the value of a large proportion of the research on attitudes. Such generalizations as are drawn are so lacking in generality that it is difficult to see their worth as a part of social psychology as science. Sometimes an investigator is careful to circumscribe his conclusions with necessary qualifications as to their generality but that is no guarantee that these limitations will be heeded by others. Although it may be important to know something about the attitudes of the college sophomore, a science of attitudes cannot be built around this member of the species. We need more studies based on samplings from a far broader universe in order to have generalizations worthy of the name. Such studies will, of course, be more expensive to execute, but would it not be better to have one

study with conclusions applicable to the generality than to have a dozen with conclusions inapplicable to the generality?

We hasten to add that the situation in the study of attitudes by scales is no worse than in other areas of psychology. The existing science of human behavior is largely the science of the behavior of sophomores. Too much research effort is expended on college students with subsequent waste of journal space devoted to speculation concerning whether the findings hold for mankind in general. There may be research problems in the attitude and opinion field which can be profitably attacked on the local campus, but it is the responsibility of the proponent of such research to show how the project can lead to generalizations which are not so restricted as to be meaningless.

When one turns to the typical single question or polling type of opinion study, one finds the attempt to sample from a universe broad enough to permit the establishment of generalizations which need not be qualified because of the possible peculiarities of the population being studied. The universe is usually defined as the population of voters or those of voting age. Sometimes a specified segment of the population is studied, not because of its availability, but because of interest in the group and a desire to compare it with the general population or with some other group. From the viewpoint of studying the opinions of adequate populations, it is fortunate that research on public opinion and its measurement got started alongside or with a polling agency which; in its efforts to predict the outcome of elections, was forced to consider a practically unrestricted population.

For nationwide or for regional surveys of either the total population or special groups, the problem of securing a representative, unbiased sample is being satisfactorily solved by the Bureau of the Census and by the Department of Agriculture's Program Surveys. Both these agencies are developing what is known as area or block sampling which must of course depend largely upon known characteristics of the total population. It has been shown that area sampling procedures will yield the correct proportional representation for the control variables, such as sex, age, rural versus urban residence, race, socioeconomic status, and geographic location, which are involved in the attempts at quota sampling by the private polling agencies. The recent paper by Hansen and Hauser (54) of the Bureau of the Census contains an excellent and convincing discussion of the advantages of area sampling.

The Research Branch of the Army's Morale Services Division had to face some peculiarly difficult problems in getting a cross section of the entire Army or, say, of infantrymen. It cannot be said that these problems were solved but, in view of practical considerations somewhat

unique to the Army situation, a reasonably good sampling job was done. Sampling from the Army was apt to be disrupted by sudden shifts of personnel, but it had one distinct advantage over sampling from the civilian population: individual refusals were nonexistent, and the factors leading to missing an individual who had been designated for inclusion in the sample were not nearly as apt to be correlated with the variables being studied. Studies done in the Bureau of the Census, as reported by Hilgard and Payne (57), and Agriculture's Program Surveys (unpublished material) supply ample evidence that those who must be reached by two or more call-backs differ systematically from those found on the first call. That refusals may be a source of bias is shown by Cantril (18), who nevertheless argues that "refusals do not greatly affect the extent to which the sample secured by poll interviewers is a representative cross section of the population" (p. 122). This may be true, but when refusals run as high as 14%, it may well be questioned.

In the survey work done in the Bureau of the Census and in the Department of Agriculture, and to a certain extent in the Army the sampling is centrally controlled. Nothing is left to the judgment of interviewers, who seldom have an appreciation of sampling fallacies. This is in marked contrast with the procedure utilized by Gallup, Roper, and those making market surveys. So far as the writer can find, these gentlemen have never indicated precisely their sampling procedure. Apparently little central control is exercised except that the interviewers are scattered widely and are told to meet certain quotas as to age, sex, color, and economic status. It is reasonable to assume that the sex and color quotas are met by all except the possibly dishonest interviewer. When one considers the unreliability of the estimates of economic status, (interviewer agreement correlation of .63) reported by Cantril (18, p. 101), one is skeptical about the quotas for this control variable. The control of geographical and rural-urban quotas is entirely satisfactory as judged from figures given by Cantril (18, p. 145) for American Institute of Public Opinion (Gallup) and National Opinion Research Center (Denver) samples.

There is evidence, however, that the sampling procedures used by these two agencies lead to bias in the direction of too many from the higher occupational brackets. This type of selection is likely to be related to an educational bias. According to data given by Cantril (18, p. 148), both agencies sample about 20% too few cases from the grade school level, about 10% too many from the high school level, and 10% too many from those with education beyond high school. Cantril admits that central control of sampling is needed to overcome these selective factors, but he is inclined to believe that such would be "too com-

plicated and costly for general use by polling organizations" (18, p. 148), which is another way of saying that quality of research is being sacrificed for quantity. Further evidence of educational bias is found in the Office of Public Opinion Research (Cantril) study on expected postwar migration, reported by Bruner (11). To find that 26% of workers in war plants have had some college education, as compared with the Census figure of about 10% and the Army figure of about 14%, suggests the operation of selective factors in the sampling procedures. We are told by Bruner that the interviewers secured workers for the sample from bus stations, parking lots, in lunch stands, and around pinball machines. It would be interesting to know just what factors do operate in the interviewer's choice of individuals for a sample.

A 1945 article by Cantril (19) is of interest in connection with the problem of sampling. He compared the responses obtained by different polls when identical questions were asked at nearly the same time by two or more of the following agencies: AIPO, OPOR, NORC, and *Fortune*. The respective N's were 3500, 1200, 2500, and 5000. The discrepancies between pairs of questions ranged from 0 to 12, with an average of 3.24 for 99 comparisons. Cantril hails this as "a highly creditable performance" considering the "expected margin of error" of "3 or 4 per cent." One notes immediately from the given N's that the discrepancies are larger than expected on the basis of mathematical error formula. What we have here are empirical discrepancies of the same order as those found in previous empirical studies, which, as will be seen in the next paragraph, led to the acceptance of 3 or 4 per cent as the "expected" margin of error. Even if the 99 differences were satisfactorily within the limits of predictable mathematical error, the result would not be infallible evidence that the polls were securing representative samples. All could be biased in the same direction—we have noted in the previous paragraph that the samples for three of these polls suffer from an educational bias.

In sampling for the purpose of predicting the outcome of elections one encounters grave difficulties. It may be relatively easy to secure an unbiased cross section of those who are eligible to vote but this might not lead to a good prediction as to what the subpopulation of actual voters will do. There are three possible sources of sampling error in straw polls: the ordinary random sampling errors, the biasing error due to selective factors in the picking of individuals by the interviewers, and the selective factors involved in voter turnout. The figures, presented by Gosnell (51), for the Gallup and Crossley 1936 polls would lead one to believe that the state by state error was too great to be attributable to random sampling errors. Likewise the analysis by Katz (64)

of the 1940 election polls indicates that the regional errors for the *Fortune* poll and the state errors for Gallup's poll were greater than expected on the basis of chance. Gallup's national sample of 40,000 would mean that the standard error would be only .25%, hence his obtained error certainly (level of confidence represented by a P of less than .0001) should have been less than 1% rather than the obtained error of 3%. Without doubt bias of some sort was involved. That Gallup feels none too sure of the randomness of his straw samples is evidenced by the fact that his reported 4% margin of error is determined not from a sampling error formula but on the basis of empirical findings: it is, according to Katz (64), his average error over a period of years.

Katz (66) has also evaluated the performances of the 1944 election polls. Again the state by state errors of Gallup were so large and consistent as to suggest inadequate sampling. Crossley's poll suffered from the same limitations. Both these polls missed the national vote by a wider margin than did Cantril's OPOR, with a sample only one tenth as large, and the *Fortune* poll conducted by Roper. The uncanny accuracy of the *Fortune* poll in 1936, 1940, and 1944 is not evidence that Roper's sampling procedures are better than those utilized by the other polls. Rather it indicates his awareness of an educational or socioeconomic, hence Republican, bias which for these three consecutive forecasts has been balanced by not correcting for the relatively light voting in the South. Roper based his final prediction on the responses to a scale of attitude toward Roosevelt, mainly because he had found that the attitude scale tended to reduce the "undecided" percentage. That such a scale technique may be hazardous is indicated by Cantril's results, reported by Katz (66): an attitude toward Dewey scale gave 4.6% more Republican straw votes than did the Roosevelt scale. Cantril's forecast was obtained by averaging the results for the two scales. If Roper were aware of the possible loading in the scale technique, he must have reasoned that this would balance some other relevant factor. The unpublicised 1944 NORC poll erred in underestimating the national Democratic vote despite an effort to construct an accurate sample by the quota method. Again, it is likely that the educational or socioeconomic bias was operating.

There has been considerable alarm in certain quarters concerning the legitimacy of applying corrections to poll results. From the scientific point of view, there can be no possible objections to corrections for known bias in sampling or other faulty techniques *providing* the reasons for, and methods of, making the corrections are placed on record. This is not to say that corrections are to be preferred to the removal of their

necessity. Corrections are apt to be too inexact, and at times involve guess work which may not be shrewd.

It should be emphasized that the failure of a polling agency to predict accurately the outcome of elections does not per se mean that its sampling techniques are inadequate for opinion studies. There are too many other factors involved; Katz (64) gives an excellent discussion of these.

Regarding the size of sample required for opinion and attitude research, it need only be said here that there are no simple, hard and fast rules. In general the nature of the problem and the type and number of breakdowns needed in the statistical analysis are the determining factors. Since the essence of a science is the establishing of relationships among its variables, breakdowns are necessary, and therefore the total N proposed for a study is not so important for judging the adequacy of the number of cases as are the N's in the required subgroups. Nonsensical as it may seem, one can find opinion studies by governmental agencies in which percentages are reported for subgroups of 50 or less; the writer has noted one percentage based on 14 cases! A consideration in deciding how many cases to use in correlational studies involving either single questions or attitude scales is the rather unstable nature of correlation coefficients. When subgroups are being compared or when interrelationships are being determined, it is well to remember that the acceptance of the null hypothesis is particularly fallacious if the number of cases is small. Differences of practical importance can too easily be brushed aside because of lack of statistical significance. For illustration, we cite the recent study of the 1942 elections by Cantril and Harding (20). It is reported that "an overwhelming proportion of the issues tested turned out to have little importance in determining the outcome of the elections." When 230 cases are broken down into six subgroups the sampling errors become so large that it is indeed difficult to reject the null hypothesis even though sizable differences exist between the population parameters.

When one is interested in the study of changes in opinion, the so-called panel technique is useful in reducing sampling errors for differences or changes. A representative group is set up as the "panel" with the idea that the members can be interviewed a number of times. The statistical advantage of the panel method is due to the subtractive correlational term in formulas for the sampling variance of differences. This technique does not, however, warrant the naive claim of Stonborough that "any change in a fixed panel is necessarily a real change and if the panel has been correctly set up to represent the universe

under study, then a change in the panel *must* mean that a similar change in the universe has taken place" (120).

In concluding this brief discussion of sampling we would like to reiterate the following three points:

1. Too many attitude studies have been made on atypical groups. A real social psychology of attitudes is in need of research on groups of greater generality than college students.
2. The single question opinion studies by the several non-governmental polling agencies, AIPO, NORC, and OPOR, have been based on samples which are frequently biased. The dependence on interviewer selection of cases is likely to be the source of sampling bias.
3. Adequate and dependable cross sections can be obtained by area or block sampling.

It is not argued that all opinion and attitude studies should be based on samples of national scope. There are many projects which can be attacked locally with reasonable expectation that the findings will possess some generality. By "locally" we do not mean the local campus, but the local or some nearby community. Nearly every fair sized community will yield samples with sufficient variation on psychological and sociological characteristics to permit the determination of correlates of opinions and attitudes. The precise degree of such correlations may vary somewhat from community to community, but the divergencies will be minute compared with those between results for college students and for samples from any community which approximates typicality. There are, of course, many special groups within a community which might be profitably studied. Such community studies would involve field work and hence would be relatively more expensive than passing out questionnaires in a college class, but much could be gained by social psychologists from actual participation in field work.

Those who wish a technical discussion of sampling should turn to statistical treatises. For a good nontechnical discussion with special reference to opinion polling, the reader is referred to Stock (18, pp. 127–142), and to Wilks (129).

Erroneous Statistical Analysis. The purpose of this section is to point out some of the errors frequently found in statistical treatment, with the hope that such errors will not be repeated in future work. Since it is not desired to single out specific studies, references will be omitted but none of the illustrations given is fictitious.

One type of error has been found repeatedly in studies of changes whether produced experimentally or otherwise. In such studies it is considered good practice to have a pretest and posttest on the same sample of individuals. If scales are used, the change in attitude is measured by the difference between the mean pretest and mean posttest score; or if single questions are used, the change is the difference

between percentages. In either case the sampling error for the difference must be determined by using the formula for the standard error of the difference which includes the correlational term. It is this term which is so frequently ignored. To argue that the omission of the *r*-term gives a conservative estimate of the sampling error is sheer nonsense. The fact is that omitting it gives a wrong estimate of error. Its omission springs from ignorance or carelessness, not from a desire to be conservative. It is true that if a difference is significant without the *r*-term it would be more significant with it. But why not use the term so as to be able not only to say that the difference is significant but also to say that it is significantly greater than some named value? Of greater consequence is the danger that the omission of this term may lead to acceptance of the null hypothesis when this hypothesis should definitely be rejected.

Another mistake is sometimes made in studying changes. An experimental and a control group are set up. If a change is significant for the experimental group but not for the control group, it is concluded that an effect has been demonstrated. The proper procedure here is to test the significance of the difference between the changes of the experimental and control groups, i.e., the net change. There are some studies in which the conclusions would be drastically altered if the net changes were evaluated.

A practice which can well be discarded is that of "explaining" a difference between means, or percentages, which is statistically insignificant. Similarly, a correlation which could easily be a chance variation from zero need not be explained or interpreted as indicating other than a lack of relationship. For example, in a study by a well-known social psychologist an r of $-.06$ was inflated into an italicised conclusion that a "direct affinity" existed between the two variables.

We have already mentioned as questionable the comparison of means based on different attitude scales.

A particularly fallacious procedure is the *blowing up* of a percentage. For example, in one study 35% of a sample of 400 said they would do a certain thing. This 35% applied to the total population of 20,000,000 led to the prediction that 7,000,000 people would so act. When one considers the sampling error of a percentage based on 400 cases, one readily sees that the correct figure could be anywhere between 6,000,000 and 8,000,000. To make matters worse there was no evidence that the sample of 400 was a good cross section of the 20,000,000.

We have already indicated that percentages based on rather small N's have been reported, and that accepting the null hypothesis is relatively more risky the smaller the N's. Despite the fact that the mathematical statisticians have provided ways for evaluating statistics based

on small samples, the writer feels that the sampling techniques in social science are seldom if ever sufficiently adequate to justify using small samples. One wonders, for example, what importance can be attached to 1431 intercorrelations based on 31 cases.

The AIPO and NORC polls of opinion are based on stratified samplings, yet we cannot find evidence that sampling errors are evaluated by formulas appropriate to the stratified situation. This is discussed briefly in Cantril's book in a technical appendix by Mosteller (18, pp. 291–292), who concludes that "we may expect no great gain from the stratification over random sampling." If this is the case, it seems likely that the controls used in stratifying are not very effective. If at all effective, the sampling errors will be less than under conditions of random sampling, hence the use of formulas applicable to the latter will overestimate the sampling errors. This will not be a serious matter when judging the sampling stability of an over-all percentage. If a group difference is significant on the basis of the random sampling formula, it would be judged more significant if the formula for stratified sampling were used, hence no harm would result from using the wrong formula. Such is not the case when a group difference is on or near the borderline of significance. The use of the correct formula for the stratified situation might lead to the conclusion that the difference is definitely significant. How significant cannot be determined without using the correct sampling error formulas.

Occasionally opinion pollers state that for their N's a difference between percentages of, say, 5% is significant at the .01 probability level. If this holds for differences in the middle range of percentages, say 47% compared with 52%, it will not hold for differences at either end of the scale of percentages. Thus, for the same N's, the difference between 88% and 92% would not be judged significant on the basis of this rule-of-thumb procedure. Actually this four point difference would be significant at the .001 probability level. The danger in such an overestimate is as before: the null hypothesis may be accepted when it should be rejected. The nomograms of Wilks (130), reproduced in Cantril's book (18), for judging the significance of differences between percentages apparently do not allow for the fact that the sampling errors of percentages decrease with the distance from 50%. Wilks' chart of confidence limits for single percentages does take this into account.

Still other mistakes in statistical treatment have been noted but they are no more frequent in the attitude-opinion field than in other areas of social psychology.

Inadequate Statistical Treatment. The writer believes that the purpose of all statistical treatment should be to make analyses and to

apply tests of significance which are as simple as possible yet suitable and adequate for the problems at hand. The researcher needs to keep in mind that statistical technique boils down to the simple task of calculating concise and convenient descriptive measures for masses of data and then drawing inferences therefrom. The descriptive terms include percentages, means and medians, standard deviations, measures of relationship or association, etc., for such groups as are involved in the breakdowns. Inferences are made about differences or relationships via appropriate sampling errors or significance tests. The reader of a report on an opinion or attitude study has the right to expect a full set of descriptive terms and specific indication of the significance tests applied.

A frequent shortcoming in the studies of attitudes by scale techniques is the omission of means, and more frequently, standard deviations. Such omissions make it difficult to compare one study with another. It is particularly disconcerting to have correlation coefficients reported without the accompanying standard deviations which are so necessary in interpreting correlations. Workers who are comparing various subgroups by attitude scales should examine the analysis of variance as a flexible technique for testing significance.

The rather persistent failure of the single question opinion researchers to give the number of cases in their over-all samples and in the subgroups involved in breakdowns does not inspire confidence in the scientific value of their data. Percentage tables or percentage bar charts are presented without any indication of the N's involved. The reader is helplessly and hopelessly dependent upon a verbal description which may or may not be buttressed with evidence as to what differences or relationships possess statistical significance, or as to how the significance was tested. In Cantril's recent book (18) there are approximately 400 sets of percentages without the basic N's. Yet this is a volume which according to its preface is "purely technical" and which "should discourage vague and unsupported criticism" of opinion polling!

In the *Public Opinion Quarterly* periodic reports of current findings of the several agencies are given in percentages for various breakdowns sans the N's so necessary for an evaluation of any hypothesis which might conceivably be checked by reference to these compilations.

In general, the statistical comparisons are made on the basis of standard errors of differences between percentages. This is adequate for a two by two table, but has some dangers for tables with more categories. The chi square technique is more generally applicable and possesses properties which make it a superior method for testing significance.

In studying relationships between opinions and other psychological or sociological variables, and between various opinions, the typical

procedure is to set up contingency tables percentaged in one direction. This may seem plausible enough, but it should be recognized that the analysis of relationships via percentages is none too precise. Since practically all relationships so studied do and will show varying degrees of association, it would be far better to use, if possible, a measure which reflects the degree of correlation. That verbal description of percentage tables may be subjective and misleading is well illustrated by two examples culled from reports of a government polling agency. One relationship is described as "intimate;" had the analyst or reporter realized that the underlying correlation was only .60, it is extremely doubtful whether that misleading term would have been employed. The phrase "closely related" is used to describe a relationship which is a mere .40 on the product moment scale, and this same phrase is applied to various degrees of relationship. At another place a percentage table which leads to a correlation of .12 is referred to as showing a relationship. It is, of course, possible for such a correlation to be statistically significant if based on a sufficient number of cases, but its practical significance is nil. In regard to difficulties encountered in substituting verbal descriptions for more precise statements of correlation, it might be pointed out that statisticians have long stressed fallacies which are inherent in interpreting relationships from percentage tables.

In this sketchy discussion of statistical issues, we have attempted to show that far too many attitude studies have been based on atypical groups, such as college students; that the opinion studies have fared better in this respect by attempting, not entirely successfully, to use cross sections of the generality of adults; and that adequate samples can be obtained by the centrally controlled area or block sampling technique. We have pointed out some observed errors in the statistical analysis of opinion-attitude studies, the most frequent mistake being the use of an incomplete formula for the sampling error of changes. As to inadequate statistical treatment we have noted the failure to report basic descriptive measures, particularly the simple matter of N's in the opinion studies. We have suggested the use of two broadly applicable techniques for testing statistical significance: analysis of variance for studies based on scales and chi square for single question opinion research. It has also been suggested that in the latter type of study, greater precision can be obtained by substituting measures of association or correlation for percentage tables which purport to indicate relationships.

VI. Study of Changes

Changes in attitudes and opinions have been, and properly should

continue to be, pertinent for those concerned with a science of attitude-opinion or with any aspect of social science wherein due consideration must be given to the influence of attitudes and opinions on the social behavior, actual or potential, of an individual or of groups. Any reasonable general hypothesis as to the origin or development of attitudes and opinions would have to place great stress on the role of learning, broadly defined as the modification of response tendencies by the impact of the entire culture which surrounds the individual during his life span. It would seem illogical to argue that attitudes and opinions are innately determined or that there are innate factors associated therewith, except that it might be hypothesized that certain very specific attitudes may depend upon so-called innate temperamental traits; that the intensity factor may be related to possible innate emotional tendencies; and that occasionally an attitude may be conditioned by one's physique. The chances for demonstrating that attitudes or opinions have innate correlates appear so small compared with the likelihood of finding cultural correlates, that it seems far wiser to concentrate research on the latter.

One method of investigating the influence of cultural factors on attitudes and opinions is through study of the conditions under which changes take place. The very fact that they do take place, whether accidentally or experimentally produced, can be taken as presumptive evidence that opinions and attitudes are learned. The problem is to determine the conditions under which changes occur or can be brought about. Ultimately this would involve a theory of learning, but whether a learning theory should precede or follow such inquiry is a debatable question. Those who are primarily interested in learning theory would argue that a consideration of such theory should lead to more fruitful hypotheses as to how changes can be produced, whereas those who are mainly concerned with attitudes could argue that they can induce changes which need to be explained by a learning theory. The writer prefers to avoid this issue.

There are, in general, two ways of studying attitude and opinion changes. One is by the use of trends based upon successive samples or cross sections from some defined population, and the other is by experiment. In typical trend studies group shifts are noted, whereas in experimental studies both individual and group changes can be observed. If significant changes take place in a well-controlled experimental study, the cause is known; but this is not always the case in trend studies—the investigator must rationalize shifts in terms of intervening events with no certainty as to which is causal.

Before the methodological issues involved in the study of changes are examined, a few general considerations should be discussed. When

scales are used, the difference between means (or the mean of individual changes) can be taken as an indication of the shift which has taken place. Before one concludes that a real shift has occurred it is necessary to apply a significance test in order to be sure that the observed shift is nonchance. Statistical significance, however, does not prove that the change has social significance, and the magnitude of the critical ratio (difference divided by its standard error) does not necessarily reflect the amount of shift. Nor does the absolute difference between means indicate amount since the numerical difference is a function of an arbitrary measuring unit. Perhaps the best way to evaluate the amount of shift is to take it relative to the standard deviation of the first score distribution, or to consider the overlap between the first and second distributions. The thing to avoid is the frequent confusion between statistical and practical significance; the former can be objectified but the later cannot. For single question opinion gauging, the shift is in terms of a change in percentages and a nonchance change may or may not be of any social consequence. Suppose that the number of cases in two samples taken at different times were such that the shift from 38% to 42% for issue A and the shift from 48% to 52% for issue B were both statistically significant. The former shift might be judged of little social significance whereas the latter, being a shift to a majority, might be regarded as of great importance, provided one did not have too many reservations because of the factors discussed in Section III. As in the case of shifts in scores on scales, the critical ratio for percentage changes is not by itself an indication of the amount of change.

The evaluation of group changes is far simpler than the evaluation of individual changes. If scales are being used, a change for an individual can be appraised in terms of the error of measurement, but this cannot be done when the single question method is used. An individual shift on a scale can be judged as beyond the realm of chance response errors, therefore likely real; there is no way of being sure that an individual shift from yes to no, or vice versa, is not due to error except in the nonexistent case of perfect reliability.

It should be unnecessary to say that for the study of changes either by scales or single questions there are decided advantages in high reliability (accuracy). Furthermore, changes take on added significance if it is first demonstrated that the attitudes or opinions of individuals on a given issue have some stability in the sense of not fluctuating from day to day or week to week without apparent cause. To know just what issues people have relatively stable opinions about is also of interest.

Trend Studies. The study of changes by the trend method is best typified by the work of Cantril (**16, 17, 18**). Samples of the adult popu-

lation are polled from time to time on the same question, and the resulting percentages are plotted as ordinates with time as the abscissa. For such ups and downs as occur an explanation is sought in terms of world or national events. If an event can be found which precedes a shift in public opinion and if it seems reasonable (without too much speculative argument) that the event would lead to the given shift, it may be rightly concluded that a causative factor has been found. It is obvious that such explanations cannot be as readily accepted as in a controlled experiment. *Ex post facto* rationalizations can sound convincing but be misleading. One fact which definitely emerges from the trend studies is that the prediction of public opinion for some future even though near date by extrapolation would indeed be hazardous.

Antedating the trend studies by polling were studies of attitudes based on successive college classes of, say, freshmen. A trend of attitude is noted and usually explained in terms of such large-scale events as the depression or the World War. Researches of this type are not as clear-cut from the sampling viewpoint as are the public opinion trend studies. In the latter the characteristics of the population are practically the same from time to time, whereas freshmen may differ from year to year. An observed "change" may reflect in toto or in part the changing characteristics of the freshman population rather than the influence of, say, the World War on attitudes.

The panel technique has already been discussed as a possible way of reducing random sampling errors in trend studies. In so far as the continued cooperation of the panel members can be obtained, and in so far as being repeatedly interviewed does not lead to changes, this method can be very sensitive for measuring shifts. The assets and liabilities of the method have been set forth in two papers by Lazarsfeld (**72, 74**). One can readily agree to the following advantages: individuals improve in ability to express their opinions; additional personal information is securable each time; the individuals who change can be questioned in an effort to determine why they shifted; and the method is less expensive for some types of studies. Carried to the extreme of reinterviewing the same people over a period of 20 or 30 years, even if feasible, the panel technique would not necessarily lead to an accurate estimate of the general trend of public opinion, where "public" is defined broadly as the adult population. An upward or downward trend for the panel might be a reflection of the ageing of its members.

An extensive use of the panel method is reported in *The People's Choice* (**75**) by Lazarsfeld, Berelson and Gaudet. Out of 3,000 persons selected in Erie County, Ohio, by area sampling, four panels of 600 each were set up in such a way as to be comparable. The members of

the main panel were interviewed monthly from May through November (post election) of 1940, while the other panels (controls) were interviewed in May and only once thereafter. Aside from determining much information about those who voted Republican as compared to those who voted Democratic, the reinterviewing of the main panel permitted not only a study of the characteristics of the changers but also an investigation of the factors associated with or leading to changes. This book is a "must" for all students of voting behavior and of methodology. Those who object to the rather small number of cases in some of the breakdowns will have noted one of the chief limitations of the technique: it is too expensive to permit adequate numbers for answering all the questions which tend to arise during the analysis of data.

The panel technique is very similar to that of testing and retesting college students as they progress from the freshman to the senior year. A number of researches have demonstrated that shifts in mean attitude scores for a variety of issues do take place as students go through college. Some importance can be attached to this knowledge, but the fundamental task is to pin down, if possible, the specific causes of the changes. It is not inconceivable that college attendance per se has little to do with the changes; a carefully matched control group of individuals not in college might show similar changes. Furthermore, the changes which take place in college are complicated by the selective factors involved in students dropping out of school. It is the writer's opinion that additional studies which purport to compare seniors with themselves as freshmen will contribute little to an understanding of the formation of attitudes and opinions.

Regardless of how small or how large the mean change in an attitude as students progress through college, one can be reasonably sure that individual shifts of considerable size have taken place. The correlations reported by Bugelski and Lester (**13**), Farnsworth (**36**), Hunter (**60**), and Newcomb (**96**) for attitude scores obtained as freshmen and as seniors are .42 or less, whereas for freshmen versus sophomore scores Corey (**28**) reports values ranging from .28 to .61 for various attitudes. Darley (**30**) found test-retest *r*'s in the .60's for autumn and winter versus spring testing on the Rundquist-Sletto scales. These correlations, particularly one of .13 for the Peterson war scale, seem to indicate a lack of stability for measured attitudes. Nothing is known about the individual fluctuation of opinions as gauged by the single question method. Research could be directed at finding the factors which lead to individual changes, providing the measuring were done by reliable scales so as to permit the singling out of individual shifts which are significantly larger than the error of measurement. The writer sees no

way of doing this when the single question method is used. If 20% shift from yes to no or from one dichotomy to the other and it is known that the measurement or response error is such that a 10% shift can be expected, there is no way of knowing which of the 20% are real shifts and which are due to chance error.

The comparison of contemporary freshmen, sophomores, juniors, and seniors has been the basis for concluding that changes have taken place as a result of college attendance; but since it is possible, as shown by Newcomb (**96**), that the attitudes of these groups may differ at the time of entrance, such a conclusion is questionable.

Another type of trend study is that of tracing the "growth" of attitudes in children. This can be done by the cross-sectional method used by Horowitz (**58**), who "tested" school children at each grade level from the kindergarten through the eighth grade, or by the longitudinal or follow-up method. It is one thing to demonstrate that attitudes show growth, but quite a different task to account for the development. The fitting of curves, as in the Horowitz study, adds little to our understanding of attitude formation. In fact, curves and mathematical equations therefor tend to mask individual differences, particularly when no indication of variations about the curve is given, and it cannnot be inferred that the growth of a given child will follow or parallel the curve based on averages unless it is established, by the longitudinal technique, that children maintain the same relative positions from year to year. It may be that the attitudes and opinions of elementary school pupils are too unstable to justify study at all, except perhaps to find out why an attitude "learned" today is unlearned next week. If particular attitudes and opinions do become relatively stable for an individual, it would be of interest to know what factors lead to the stabilization. If stabilization never occurs, the chase after causes may become a will-o'-the-wisp affair.

The study of trends in public opinion should contribute to an understanding of the factors which affect opinions, and should prove of value to the social scientist who is interested in social and political developments. The trend method does not admit the elements of control possible in the experimental approach, and furthermore the results are contingent on perspicacity in choosing the issues about which opinions are apt to be changed by events, usually unforeseeable. Yet in a sense, the use of trends is the use of the nation as a laboratory, hence no loss in reality is involved as is sometimes the case in experimental studies. That trend studies coupled with breakdowns can yield enlightening results is definitely suggested by Lazarsfeld's (**71**) study of changes in opinion during the debate and discussion concerning the appointment

of Hugo Black to the Supreme Court. At first the line-up was political, later it was religious.

Changes by Experimentation. The literature contains some thirty reports of experimental studies of changes in attitudes or opinions. Most of these studies have involved measurement by scales. Some researchers have been primarily interested in attitudes per se and changes which can be produced, while others have been more concerned with the effectiveness of various media as possible molders of opinions or attitudes. Studies have been carried out on college, high school, and elementary school cases, hence the conclusions may not be applicable to other groups. One would expect the results obtained at the school level to hold generally for groups of similar level, whereas one cannot claim that findings based on college students can be generalized to other adult groups.

The experimental method has been utilized to investigate the effect of the following: propaganda, written and orally delivered in person or via radio; editorials and informative literature; classroom teaching of various sorts; group discussions; knowledge of expert or of majority opinion; and movies which might tend to shift attitudes. A little reflection will make it obvious that experiments based on different media cannot achieve the same degree of reality; if the subjects recognize the setup as artificial, the results may not be too dependable.

The essentials of a good experiment designed to determine the effect of a given experience on attitudes will be familiar to most readers although not all researchers using this approach have conformed thereto. An experimental group, for which an experience is provided, and a control group, which presumably is subject to the same influences as the experimental group except for the particular experience, are set up by random drawings of individuals from one defined population and by chance assignment to two groups. Greater precision can be attained by pairing of individuals for the groups, the matching to be done preferably on the basis of initial or pretest attitude scores or on the basis of control variables such as age, grade, sex, socioeconomic status, etc. The writer has discussed elsewhere the advantages of the matching technique (85). In practice it is not always feasible to use either the random or the matching method for forming groups. If one must resort to intact school groups, the aim is to choose two groups (each of which may be made up of subgroups) which are as similar as possible. Strictly speaking, this requires more complicated statistical treatment of results, but it is doubtful whether the use of the ordinary random error formulas would be seriously misleading.

Both the experimental and control groups are given a pretest, say

on Form A; then after the experimental group has been exposed to the given stimulation, both groups are retested on form B. If only one form of a scale is available, or if the measurement is by single questions, the pre- and posttest must be based on the same form or the same questions. When the pretest and posttest data are in hand, the particular statistical treatment utilized varies. Suffice it to say here that it is inadequate and inefficient to test the significance of the difference between the two posttest means even though the difference between the two pretest means is also tested. Nor is it adequate to determine the significance of the difference between the pre- and posttest means (or the mean of the changes) for each group and conclude that an effect has been demonstrated if the experimental group shows a significant change and the control group does not. The proper thing to do is to test the significance of the net change, that is the difference between the changes for the two groups. This makes due allowance for any difference in the initial or pretest standing of the two groups and also brings out what is desired—the net difference in change reflects the influence of the experience interposed for the experimental group over and above the influence of other, and uncontrolled, factors common to the two groups during the interval. In computing the sampling error of the change for each group (needed in calculating the sampling error of the difference between changes for the two groups), one should avoid the frequent mistake of not including the correlational term.

If single questions are used, the changes will be in terms of percentages instead of means, but the general pattern of proper statistical treatment will be the same as just suggested.

There are times when a pretest is not at all feasible. This has been the case in some of the experimental work of the Research Branch of the Army's Morale Services Division on the efficacy of orientation films. An experimental group is shown a film, then tested at the same time that a carefully chosen control group is tested. Any significant difference between means (or percentages) for the two groups on this one testing (corresponds to the posttest) is taken as indicating that the film has had an effect. Obiously this design is not to be preferred over that permitting a pretest, but dependable results can nevertheless be secured providing the individuals can be selected either by strictly random methods or by pairing or matching in making up the two groups. If matched groups are used, the sampling error of the difference must include a correlational allowance for this fact. Some have been apprehensive about this single test experimental method because of the possibility of the groups not having the same initial standing. It should be noted, however, that the random formation of groups will not lead to

initial differences larger than those attributable to chance, and that such chance differences will not produce nonchance differences in the end results. Pairing or matching of individuals on available control variables should help to allay apprehension concerning such an experimental design. Actually the situation is precisely the same as drawing samples, by the random or by the stratified method, from two populations; a population of those who can be provided with an experience like seeing a movie and a population of those who have not had this experience. The sample statistics are then used to determine the likelihood that the population parameters are different. The precision of the single test scheme will, in general, be less (sometimes much less) than that of the pretest-posttest setup, but this loss of precision can be overcome by taking more cases. For further discussion of these points the reader is referred to the previously cited article of the writer (85) wherein will also be found the appropriate statistical formulas for evaluating the difference between groups which have been formed either by individual pairing or by matching groups on control variables.

A quasi-experimental approach has been used in a number of attempts to demonstrate that changes in attitude and opinion occur as a result of taking certain courses of study. In the absence of a control group, one does not know whether the course or something else has produced such changes as occur.

In connection with the experimental production of changes the question is sometimes raised as to what types of persons show the shifts. This is a commendable quest but some who have taken it up have fallen astride the regression fallacy. If a group is given one form of an attitude test today and another form tomorrow, it will be found that those who were initially more favorable will be on the average somewhat less favorable on the second testing, whereas those who were more unfavorable initially will appear to become less unfavorable. Now, if in an experimental situation the mean for a group has shifted, say in the direction of a more favorable attitude, it will be found that those initially near average will have changed about as much as the total group, those initially favorable may have shifted very little or none at all, and those initially unfavorable will have changed much more in the direction of favorableness than has the group as a whole. This is the well-known regression phenomenon. The fallacy is that of concluding that the experience has had a differential influence according to an individual's initial standing. Actually this apparent differential effect may be nothing more than the regressive effect due to the ever present errors of measurement, which are, as we have seen, fairly sizable for attitude scales. The regressive effect of measurement errors will also tend to

produce a correlation between initial standing and change. This is the same regression phenomenon which the writer has discussed in more detail elsewhere in connection with the study of IQ changes (84). It is not surprising to find instances of the regression fallacy in the attitude literature.

That attitude changes can be brought about by various means has been amply demonstrated by the experimental studies. It has also been shown that induced changes have persisted, though with some loss, for as long as six months. The writer is inclined to believe that, considering the possibilities for determining by the experimental method the formative causes of attitudes, only the surface has been scratched. Opinions could be studied likewise, but much more profitably if the single question were replaced by a scale. The possibilities are nearly unlimited, and the rewards in terms of understanding how attitudes and opinions come to be formed or changed are doubtless great. This is one way in which experiments can be set up with controls sufficiently adequate to guarantee the dependability of results, provided reliable and valid measures of socially important attitudes can be constructed.

VII. Correlations and Interrelationships

It can be argued that the fundamental task of any science is to discover and formulate the relationships between certain of its variables. The variates of a science may be thought of as those which are commonly accepted as such, plus those which have been newly isolated and those as yet undiscovered. Interrelations are sought as a part of scientific descriptions for the purpose of making predictions or providing "explanations." Thus the determiners of those variables known as attitudes and opinions are sought in terms of sociological, psychological, and sometimes physiological correlates; and the degree of intercorrelation among attitudes and opinions is of interest as a part of the description.

In the study of relationships two necessary considerations must be kept distinct. One pertains to statistical significance, the other to the degree of relationship. Before we can conclude that there is an association between two variables, some appropriate test of significance must be applied, but this is not sufficient even though a large number of researches on public opinion do stop at this point. A measure of the *degree* of association or correlation provides a more precise and objective description and permits a comparison of various relationships. High statistical significance does not necessarily mean a high relationship, nor does a low, though acceptable, level of significance always indicate a low degree of association.

Correlational analysis is most meaningful when classification on the variables being correlated is reliable. The attitude and opinion literature contains far too many instances of analyses based on measures of low reliability. In studying relationships it is also important to use groups for which the variation under consideration is unrestricted. If one wished to know the degree of correlation between height and weight for young adult males, one would not determine this on the football squad, yet such an absurdity occurs again and again in the study of correlates of attitudes. Just why one should allow a study of covariation to be limited by a restriction of the very variation being investigated is difficult to understand. This is not to say that research should be based on a group which is heterogeneous with respect to age, nor is it to imply that relationships based on groups so selected as to hold constant certain relevant variables are not of importance. The fact of the matter is that investigators using attitude scales have seldom chosen their groups with any serious thought as to the meaning of relationships which might be found. The public opinion pollers are not subject to this criticism.

Another question concerns what correlates and intercorrelates one should attempt to establish. One would naturally suppose that the answer to this would be found in the hypothesis stage of a study. Sometimes this is the case in attitude-opinion research, but more frequently the hypothesis step seems to have been displaced by a "let's see what this correlates with" or "wonder whether it will differentiate between this and that group." If anyone wishes to argue that such a criticism is unfair, let him then explain the abundance of negative findings reported in the literature. Surely carefully conceived hypotheses would not lead to such a dearth of positive results. For instance, why would any psychologist expect to find that attitudes of students toward Sunday observance, the church, and God would be related to receiving aid from the NYA? Or take a recent report (76) based on a mammoth questionnaire of 3000 items, laboriously filled in by 409 individuals. Some of the 3000 items were scored as a "conservatism-radicalism" scale, then it was ascertained that 1076 of 2600 items were related to conservatism-radicalism. Since by the statistical level of significance used, 5% or 130 of the 2600 items would be "significant" by chance, we infer that the author was correct in 946 and incorrect in 1654 of his hypotheses, that is if he started with hypotheses.

The Quest for Correlates. In the typical scale study the correlates of attitudes are determined either by breakdowns or, in case of graduated variables, by computing correlation coefficients whereas the single question opinion studies nearly always depend upon breakdowns. When breakdowns are used, the relationships are seldom stated very precisely.

In some scale studies, analyses are also carried out on the basis of the items. If the original scale is good enough to justify analysis of total scores, this item work is somewhat superfluous, especially when one considers the likely low reliability for responses to items. We have already stressed the importance of stating relationships in terms of degree of correlation. Certainly, if the task is to account for the variance in attitude by finding possible determiners among the so-called more fundamental variates, measures of correlation will be needed.

The search for the correlates of attitudes and opinions has been among both sociological and psychological variables. Typical breakdowns are for age, sex, rural versus urban residence, geographical region, racial or national origin, religious affiliation, education, socioeconomic or income status, political party, size of family, occupation, club memberships, etc. Percentages or means for the various subgroups are compared. It has been found that numerous attitudes and opinions do show differences for many of these breakdowns. Sometimes these differences are interpreted as throwing light on the origin of individual differences in attitudes or as indicating some of the determiners of opinions; at other times the emphasis is on the comparison of groups per se. Even if one were to grant that some combination of factors upon which the breakdowns are made could be utilized in "explaining" attitudes and opinions, it would be nearly impossible by this means of analysis to specify what proportion of the variation in attitudes had been so explained. But we can be reasonably sure that the residual variation left unexplained would be found upon further analysis to be surprisingly large, perhaps so large that it would have to be concluded that only a small percentage of the variance had been accounted for. The real challenge is the residual variation.

Correlates have been sought among other variables, such as specific information about issues, intelligence, personality characteristics, family and personal background factors, types of experience, etc. Parent-child correlations have been determined for some attitudes. Each attitude is apt to be related to a large number of interrelated and independent variables, including those mentioned in this and the last paragraph, none of which singly or in combination will be found to produce sizable correlations with the given attitude. It is the opinion of the writer that the quest for correlates among sociological and psychological variables, especially by the dragnet of many-item questionnaires, will continue to be rather unprofitable as a means for discovering the determiners of attitudes and opinions. Furthermore, attitudes and opinions which are relatively unstable are not apt to yield correlations with background factors. It may well be that some attitudes and opinions are too com-

plex in origin to permit the finding of any very clear-cut determiners. If so, we may have to be content with the establishment of rough tendencies by the breakdown and correlational techniques. Such correlations as have been reported have been attenuated by measurement errors, by the use of selected groups with restricted variation, and possibly by the presence of more than one dimension in the measured attitudes and opinions.

Interrelationships. Those who have studied interrelationships among attitudes and/or opinions have been about as indifferent to hypotheses as those who have searched for correlates. In general, interrelations of attitudes have been determined on the basis of scales and on restricted, usually college, groups. The covariations are usually expressed in terms of correlation coefficients in studies based on scales, and as percentaged tables for cross tabulations in single question studies. The relatively low reliabilities for single questions become quite disturbing in determining the covariation of opinions. The reason for ascertaining correlations among attitudes is to learn what attitudes go together, or tend to form clusters. In the opinion gauging field, the term "patterns of opinion" has been used as somewhat analogous to clusters. Once a table of intercorrelations has been prepared, it is relatively easy to pick out the attitudes and opinions which tend to cluster together. This can be done roughly by inspection or more exactly by means of factor analysis. The patterns of opinion spoken of by Cantril (**18**, pp. 185–190), however, seem ill-defined and vague to the writer. Perhaps we have a bias in favor of correlation coefficients for this sort of thing.

The establishment and study of interrelationships of attitudes or opinions soon lead to two closely related questions: How many dimensions are needed to describe the attitudes and opinions of individuals? To what extent are attitudes general as opposed to specific?

The sense in which general and specific have been used in describing attitudes needs some clarification. Some have had in mind the connotation used by Spearman in connection with his two-factor theory of intelligence; but, instead of thinking of a general factor running through all attitudes, they have regarded a "general" attitude as restricted to a group of attitudes, e.g., conservatism-radicalism with respect to several social issues. Accordingly it would be postulated that an individual's score on a particular test depends in part on his general attitude and in part upon a factor unique or specific to the particular scale. It would be foolish to expect to find a general factor running through the aggregate of measured attitudes. This would imply that all attitudes are correlated, which certainly is not the case; hence the term "general" must be thought of as applicable to a family of attitudes such as

conservatism-radicalism, morale, internationalism, etc. At this point we are not concerned with how many general or broad group factors might be required to explain the intercorrelations within a family of attitudes.

A second sense in which general and specific have been used in attitude research pertains to the form of statement utilized in attitude scales. Obviously a statement can be written in general form, e.g., "the Church is a valuable institution," or more specifically, "the Church is a solace to widows." The chief problem in this case is whether a scale made up of general statements and one made up of specific statements tend to tap the same attitude. The meaning of general and specific has, in this usage, a connotation entirely different from that of Spearman.

It is not our purpose to discuss the pros and cons of the controversy concerning whether attitudes are general or specific. The matter can best be settled by further research, and there is no need to hamper observation by postulating that personality is too integrated to permit specificity. Moreover the terms "generality" and "specificity" are relative to the situation—there are degrees of generality and degrees of specificity.

This whole issue, however, does have an important bearing on measurement theory. If someone proposes to measure a new attitude and assembles a set of statements for the purpose, he must show that the retained statements do tap something in common. The fundamental requisite of uni-dimensionality for attitude measurement means that the statements must have only one thing in common. It may be that general statements will be best for the purpose, but for some attitudes specifically phrased statements may be more adequate; even the general statements will always be found by statistical analysis to have some specificity. A second implication for measurement is that certain families of attitudes may not involve a general factor to the exclusion of group factors. The meaning of factors is never too clear even when they are derived by the neat mathematical devices known as factor analysis techniques. The general factor is frequently the most difficult to define—in fact, some factor analysts would not agree that such a factor exists. One school would cause it to vanish by rotations, another would keep it, but all agree that any of the factor methods will answer the question as to whether one or more than one factor is needed to explain the intercorrelations.

There is at least one family of attitudes whose meaning would be clarified if studied by factorial techniques. The attitude or attitudes in question have been the subject of a large number of publications, and many scales have been proposed as measuring instruments, but it is not

always clear just what is being measured. Even though a given investigator knows what is involved in his measuring instrument, the attached label "conservatism-radicalism" is no guarantee that he has used the label in the same sense as that intended by any of his contemporaries. Not only do various scales constructed to measure "conservatism-radicalism" seem, by inspection, to be getting at different things, but even within a given scale there may be a mixture of kinds of "conservatism-radicalism." Furthermore, several of the Thurstone scales designed to measure rather specific attitudes are interpreted by some as measures of "conservatism-radicalism;" at times, four or five of the scales are used in a study to characterize people as conservative or as radical even though the intercorrelations among these scales may be low positive, zero, or low negative.

Apparently there may be different kinds of conservatism and radicalism, or this attitude may have many facets. Until such time as it has been established that radicalism-conservatism is a unity, it would be well for researchers to alter their labels to read "conservatism with respect to ——." The existing confusion might be alleviated by a thorough factor analysis of the several attitudes which might be considered as candidates for the conservatism-radicalism label. First, reliable scales should be constructed for each of the seemingly different aspects. Each of these scales should be as nearly as possible a pure measure of conservatism-radicalism on a given issue, i.e., no scale should be a conglomeration. Thus there would be assembled a battery of scales for measuring reactionism or conservatism versus liberalism or radicalism on a number of issues. These scales should be administered to a cross section of adults, with age controlled, for the purpose of calculating the intercorrelations. To insure some stability in the factor loadings, upwards of 400 cases should be used and the correlations should be computed by the product moment method. In spite of possible divergent interpretations of the factor analysis results, such a study would do much to clarify the meaning of this attitude cluster.

The judicious use of factor analysis might lead to enlightenment in other attitude and opinion areas. Little, however, would be gained by indiscriminate factor analyzing of every available table of intercorrelations, nor would a giant factor analysis involving all the available scales add much to our understanding of attitudes. It is very likely that many of the scales have little or nothing in common—factor analysis is designed to explain covariation, not its absence.

Unfortunately, the literature is already revealing some of the questionable benefits of factor analysis. To illustrate what this new statistical tool leads to when applied to the attitude field without sufficient

foresight, we shall briefly review Ferguson's recent series of papers, the first published in 1939 and the seventh in 1944.

Ferguson began with the intercorrelations of the Thurstone scales for attitude toward: (1) God, (2) evolution, (3) birth control, (4) war, (5) capital punishment, (6) treatment of criminals, (7) patriotism, (8) communism, (9) law, and (10) censorship. In the first two factor analyses carried out (**39, 41**), the last four scales did not behave in a rational manner, so they were purged. The remaining six scales yielded two factors: one was defined in terms of scales 1, 2, and 3 and named "religionism"; and the other, called "humanitarianism," involved scales 4, 5, and 6. Then in a later paper (**43**), by methods which would be questioned by expert factor analysts, Ferguson succeeded in "isolating" a third factor, defined by the last four scales and christened "nationalism." He has dubbed these three factors "the primary social attitudes."

One of his 1944 papers (**45**) reports a repeat analysis based on a new and larger sample of college students; since the same technique yielded the same factor alignment, we can at least be sure that "the primary social attitudes" are real in the sampling sense. The reality of these "primaries" is not readily grasped, however, when one considers the extent to which the defining scales are intercorrelated. For the first three, the basis for "religionism," the coefficients are .20, .15, and .46; for the three which are used to define "humanitarianism," .47, .24, and .39; for the last four, from which "nationalism" was elevated, .19, .24, .33, .32, .26, and .27. No further remarks are needed for those who have an appreciation of the fact that such low correlations indicate that the defining variables have little in common. Ferguson proceeded to set up scales for measuring the primaries (**40, 43. 45**), and for some of the measures he reports the impossible: higher "validity" than reliability. There is no evidence that the derived scales are unitary measures.

After "the primary social attitudes" had been discovered, isolated, and measured, Ferguson's next step was to search for their correlates (**46**); then—this is the capstone—he determined the correlations of "the primary social attitudes" with "national morale" (**44**). The lack of significant correlations in these two papers doubtless reflects a want of starting hypotheses.

Actually, granting that primary social attitudes exist, one would not attempt their discovery by way of a few selected attitudes. One would need to include *all* measured attitudes, with intercorrelations based on unselected cases rather than on college students. Even then there would emerge only the primary components underlying the attitudes for which there are measures, and these primaries might hold only for the population being sampled. Factor analysis may be a useful tool for certain problems, but one is skeptical as to its value in determining components in such a broad and diverse field as that of opinions and attitudes.

Perhaps progress in establishing the correlates of attitudes (opinions should be measured by attitude scales) can best be made by first ascertaining what attitude variables are judged by psychologists and sociologists to be of the greatest social and scientific importance. Then a concentrated effort should be made to find all possible correlates of each such attitude, or cluster of attitudes. Separate research programs could be set up for each attitude or attitude cluster. Each would involve an extensive program of research, but a lot of the hit-and-miss attempts could be avoided by giving due thought to the hypothesis stage. A large proportion of the false hypotheses might be eliminated in thorough critical discussion by a group of researchers with diverse theoretical backgrounds. If three or four such programs were carried out, we should be in a position to judge the fruitfulness of an extensive approach. It seems reasonable that a few such studies would contribute much more than the same amount of effort spread piecemeal over a variety of attitudes, some of which are of only passing or personal interest, and the study of which is too limited to provide anything more than qualified conclusions.

VIII. Studies of Morale

The national disunity in 1941 concerning our part in World War II, plus the continued lack of unanimity after we were plunged into war, was of such grave concern that there was rush to diagnose and prescribe a cure for the apparently low morale of the nation. Morale seminars were organized, governmental agencies called in social scientists to study civilian morale, and the War Department set up an organization to carry out research on the morale of troops. It would appear that the social psychologists of the country had only a few facts but much in the way of so-called principles to offer. Facts were lacking because so little direct research had been done on morale; nevertheless by 1942 a 463 page book on *Civilian Morale* (126) had been compiled by social psychologists. This purported to be scientific, but it is our opinion that the available factual data and legitimate speculations could have been summarized in less than 100 pages.

In the absence of a sufficient body of scientifically established facts and principles, the next best offerings of the social scientists were methods and tools for studying morale. The principal techniques available were those of attitude scaling and opinion polling. It is the purpose of this section to evaluate the chief efforts at defining and measuring morale. We shall proceed on the defensible assumption that scientific study of morale is impossible without some type of measuring stick; that is, we must be able to classify people as to their morale in order to discover what factors affect it.

What is perhaps the first serious attempt to measure and study morale is reported in the 1936 book by Rundquist and Sletto, *Personality and the Depression* (**112**). For these authors morale is "an exceedingly generalized trait," connoting "zeal, hope, confidence in oneself, and in what the future will bring . . . one's ability to cope with the future" (**112**, p. 201). Symptoms of low morale are: distrust of people, belief that the future is black, and that life is not worth living. They report a split-half reliability of .80, and test-retest (60 day interval) coefficients of .72 for men and .61 for women, for a scale of 20 items constructed and scored by the Likert technique. Evidence that the scale possesses some validity was obtained by comparing various groups. Their scale correlated with Hall's (**53**) occupational morale scale to the extent of .41 and .48. The Rundquist-Sletto scale has since been pitted by D. C. Miller (**87**, **88**, **89**) against the items of a 52 page questionnaire. This is another example of the dragnet method (cf. the 3000 item schedule mentioned in the preceding section).

Little need be said here concerning the Whisler and Remmers (**128**) three-way scale for measuring morale except that the reliabilities were only .55, .56, and .36, and that correlations with the Chant-Myers (**22**) optimism-pessimism scale were near zero. It is of some interest to note that the intercorrelations of the three parts of their scale were .57, .16, and .20. Whisler and Remmers' success in fitting perfectly a third degree parabola to four observation points does not, of course, add even an iota to scientific knowledge. This merely demonstrates a known mathematical fact in curve fitting.

The work on morale in industry also affords an indication of the measuring techniques available at the onset of the wartime crisis in morale. A typical procedure is that used by Kolstad (**69**) in a study of employee morale and attitude in a department store. A 10 item scale was developed for measuring that aspect of morale known as job satisfaction; then an attempt was made to find what specific situations or "gripes" were related to this general job satisfaction. The reader is referred to Blankenship (**5**) and Hull and Kolstad (**59**) for adequate discussions of methods of studying employee morale. For a good review of the morale literature which appeared prior to 1941 the reader should turn to Child (**23**). Here we shall consider briefly the studies initiated as a result of the war situation.

Miller (**90**) set himself the task of measuring "national morale," which he defines as pertaining "to all factors in the individual's life that bring about his energetic participation in the tasks which most effectively secure the national goals." He began by hypothesizing that national morale has five components, then deduced a total of 48 items from his hypotheses, and found that the 18 items having the highest internal consistencies included items deduced from all five postulated components. From this he concluded that his hypothesis concerning the components of national morale had been verified. Ac-

tually this does not constitute a verification, but does show that what one gets out of the statistical mill depends upon what was put into the hopper. Miller's resulting scale may have anywhere from one to perhaps six or seven components. The point is that he has not demonstrated by appropriate statistical treatment how many components are involved in his scale. Reliability coefficients running from .70 to .88 are reported in a second paper (**91**), which also gives a biserial correlation of .51 for low national morale and affiliation in organizations protesting the United States' entry into the war. Since this "validity" coefficient was based on a 2% dichotomy it must be regarded with some skepticism. Furthermore, certain statements in the scale are specifically concerned with interventionism, hence a relationship should not be surprising.

In this same paper, on the basis of differences found between samples from Washington State, Oklahoma A. and M., New Hampshire, Smith University (colored), and an unspecified college in Indiana, Miller concludes that there are regional differences in the national morale of college students. Maybe so, but no evidence is given that the five institutions are representative of their regions, nor that random samples were drawn from these five. The fact that he includes an item analysis comparison of the five groups indicates a lack of confidence in total scores or an admission that there may be 18 components in national morale. In a third paper (**92**) Miller reports a change to higher morale among Washington State students as a result of Pearl Harbor. The change is on the borderline of statistical significance. Again he turns to item analysis and discovers an upward shift for some items, a downward shift for others. This creates faith in neither the meaning of total scale scores nor what can be deduced from item analysis. But if one recalls the difficulties which plague the single question method, the inconsistencies for single items are not surprising. Diggs, Hanger, and Mull (**31**) have reported a correlation, based on college students, of .14 between the Miller and the Rundquist-Sletto scales. Evidently "national morale" is something different from morale as measured by the latter scale.

A different procedure has been followed by Harding (**55**) in constructing a morale scale. He started with 48 items based on a definition of morale similar to Miller's, found the intercorrelations among these items, and thence several clusters, four of which were used as the basis for a new edition of 59 items. These clusters have to do with confidence, tolerance, realism, and idealism. Then 20 of the 59 items were retained because they differentiated between a "high" morale group of 44 "distinguished social scientists and men in public life, each of whom is devoting a considerable part of his time to the furtherance of national morale," and a "low" morale group composed of 14 "surly" young men, mostly under 20 years of age, from a settlement house, and 32 men loafers on the Boston Common. That the "high" and "low" groups prob-

ably differed on a host of uncontrolled but relevant factors seems obvious. It is of interest to note that seven of the eight Rundquist-Sletto items tried on these two groups failed to show a difference. We infer from this that the Harding scale may have little in common with that of Rundquist and Sletto.

In an effort to find some of the personality-social correlates of morale, Sanford and Conrad (**113**) gave the Harding scale to 100 men and 173 women in a course in Mental Deficiency (what universe this sample represents is not specified). The report is based mainly on the results for men and presents the relationship of morale to 12 of 53 items in a questionnaire which the subjects also filled in. The authors do not reveal their reason for choosing to report on these particular items; if these 12 were the only ones which yielded relationships approaching statistical significance, we obviously have a hazardous capitalization on chance. No significance tests are reported; our casual calculations indicate borderline or less than borderline (5% level of confidence) significance. For example, "one of the most discriminating questions" concerned the number of children desired by the subjects. Calculation shows that the correlation between this and morale is .06 for 88 cases. By no stretch of the imagination or by no amount of statistical juggling can so small a correlation be considered significantly different from zero; yet these authors state that "it is clear that the high-morale men tend definitely to differ from the low-morale group in their expressed desire for children "(**113**, p. 13). This inflation of a correlation of .06 is accomplished by considering only the 9 high and 8 low morale men and ignoring the question of statistical significance. Sanford and Conrad found a correlation of .23 between the "confidence" subscale and the remainder of the Harding scale. Apparently the confidence aspect of morale has so little in common with the tolerance, realism, and idealism aspects that the meaning of a total score on the Harding scale is questionable.

A "life satisfaction index" of morale has been proposed by Goodwin Watson (**127**) This index is made up of four parts: liking of activities, liking of people, toughmindedness, and freedom from unhappy symptoms. These components intercorrelate to the extent of .54, .19 and four −.04's. The correlation of .54, between the two last named parts, could easily be spurious to the extent of .50 since toughmindedness is measured by only two items, one of which is nearly exactly repeated in the "freedom" component and the second of which is very similar to a "freedom" item. The "freedom" part contains eight items. Watson's index is based on markedly disparate parts, and as such is unequivocally meaningless.

Cronbach has defined morale as "the tendency of the individual to predict realistically the hardships or lack of hardships he will face in the future" (**29**, p. 12). He constructed a 50 item test for measuring "general optimism" and a 20 item test for "personal optimism," both having to

do with the war. Apparently a number of items should have been eliminated from the scoring because of low internal consistency. The reliabilities were .78 and .56 for the general and personal scales for 980 high school boys; and .78 and .68 respectively for 1069 high school girls. The correlation between the two scales was .45 for boys and .40 for girls. A rough cluster analysis of some of the intercorrelations among the general optimism items led to the conclusion that morale, as here measured, is not a unitary variable, but may have three components: morale with respect to military progress of the war, civilian sacrifices and hardships, and personal freedom. A search for background correlates of morale proved futile.

On the assumption that one component of morale is "confidence or optimism," Conrad and Sanford (**25**) have prepared a 10 item scale to measure "optimism with regard to the military prospects for an early, easy victory," a 14 item scale to measure "optimism concerning the consequences of the war," and a 24 item scale "designed to measure general or personal optimism." To call these collections of items "scales" seems unwarranted since nothing is said concerning how the items were selected or whether any were eliminated. Apparently the old a priori method of the 1920's was used. Perhaps this partially accounts for the low form-versus-form reliability coefficients: .45, .62, and .60 respectively for the three scales and .62 for the first two scales combined. The corresponding split-half reliabilities were .49, .68, .70, and .65. A second article (**26**) was devoted to responses to the individual items even though it had been admitted that reliability of the single items is likely to be less than .25 (**25**, p. 302). These scales were given to University of California students, most of whom were taking a summer (1942) course in Mental Deficiency. There are several points in this study which should be examined.

First, we note a tendency to assign absolute meaning to the mid-scale point as indicating a neutral position. Thus the group of students is characterized as being pessimistic about the consequences of the war and yet as possessing general and personal optimism. Perhaps this is true, but surely the authors are well enough versed in psychometrics to know that such absolutes are highly questionable. Changing the wording or form of the items can easily lead to a shift in means. A similar interpretation of arbitrary score units as having absolute meaning is found in their conclusion that the obtained standard deviations can be taken as evidence for lack of homogeneity or unanimity. In the article on individual items, it is again concluded that unanimity of response is lacking as judged by the spread of responses over as many as four of the five possible points for an item. The variance of this spread is about 1.00; now if we accept their own estimate of item reliability (or unreliability) as being less than .25, it can be said that at least 75%, perhaps 80%, of the lack of unanimity is due to response or measurement

errors. If we reversed the Brown-Spearman formula, we would find .078 as the item reliability for the military optimism items and .115 for the "consequences" items. Maybe as much as 90% of the lack of unanimity is a function of response errors!

Some of the correlations reported by Conrad and Sanford are of interest. Military optimism and "consequences of the war" optimism correlate .33, or .57 corrected for attenuation, thus showing that the two scales have something in common. The correlation between military optimism and the combined scale (sum of the two) is .73; whereas "consequences" correlates .88 with the combined scales. The difference between these two coefficients was explained (**25**, p. 304) on the basis of the larger standard deviation for scores on "consequences." A logical and more pointed explanation is the difference in degree of spuriousness when scores on 10 items are correlated with scores based on these 10 plus 14 others, as compared with that when scores on 14 items are correlated with scores on these 14 plus 10 others. A correlation of .60 is reported for the combined "war optimism" scale with D. C. Miller's national morale scale. If this were corrected for attenuation it would become about .85, thus demonstrating that two of the many morale scales so far discussed really tap the same thing. What that thing is is not made clearer when one notes that an individual was considered to have low morale if he believed in 1942 that "Germany will probably bomb" our industrial centers in the East. Such an opinion would be tantamount to *high* morale in the sense of having confidence in leaders, since a desperate effort was being made to inculcate that very belief. One other finding of Conrad and Sanford should be noted: they estimate that the average intercorrelation among the military optimism items is .08, among the "consequences of the war" items, .12, and among the personal or general optimism items, .07. Evidently these aspects of morale have little in common, so the meaning of total scores is enigmatic.

Another major research on civilian morale appears as a 50 page report in Cantril's *Gauging Public Opinion* (**18**). This study reveals an awareness that morale is a complexity of attitudes. Twenty-five questions were carefully formulated to tap 12 components or aspects of morale deduced from a study of previous research. These questions were included in a ballot which was administered in November, 1941, by interviewers to a sample of 2543 adults, stratified for section of the country, city by size versus farm residence, economic status, sex, and age. It will be noted that this is the first published study of morale based on an adequate sample of the people. In order to determine whether the postulated components were independent of each other, the intercorrelations among 23 questions were computed, and the resulting table factor analyzed by the centroid method. Two factors were found sufficient to account for the correlations. The first centroid factor was identified as interventionism, i.e., high morale was associated with

interventionism. If the reader is surprised at this, he should recall that what comes out of factor analysis depends upon what has been put into the mill.

The second centroid factor was called "degree of sophistication," but the opposite end of the axis seems to involve satisfaction with group progress and confidence in the leader, the news, and the armed forces. Apparently rotation of axes would not clarify the meaning of the two factors. Nor is the meaning of the second factor made more intelligible by the discussion on page 241 of "patterns" of intercorrelations. It is said that "The questions measuring *Awareness of the objective* and *Amount of information* are highly correlated with each other and moderately correlated with the Agreement and Determination questions." The last group of questions has to do with interventionism. What is meant by "highly" correlated? Coefficients of .30, .40, and .59. What is meant by "moderately" correlated? Eighteen correlations with a median value of .17, but including coefficients of .30, .33, .36, and .41. The above quotation is followed by "On the other hand, *Confidence in the leader*, *Confidence in the armed forces*, *Confidence in the news*, and *Satisfaction with group progress* are also highly correlated with each other and moderately correlated with Agreement and Determination." What is meant this time by "highly" correlated? Twenty-eight coefficients ranging from .66 down to −.13, with a median value of .24. And what does "moderately" correlated mean? Forty-eight coefficients running from −.18 up to .63 with a median value of .25.

The 19 questions yielding the highest loadings on the first factor were scored as a morale scale. Obviously such a scale is not unidimensional, but its saturation with interventionism is sufficient to produce differences along expected isolationist-interventionist lines. The scale reliability is not reported, but from the intercorrelations we infer that it would be between .80 and .85. With morale so closely identified with interventionism, it was of course necessary to revise the ballot for use after Pearl Harbor. The new ballot contained 27 questions designed to get at 16 postulated components of morale, and four questions designed to measure participation in war activities, "as a criterion of morale against which questions and components could be validated" (p. 243). This new ballot was administered to a cross section of 2539 adults in March, 1942. Most of the 16 components did differentiate between high and low participation individuals; we are not, however, provided with any numerical values by which to judge *how* important a component was. That the battery did not contain 16 statistically independent components is revealed by a factor analysis of the intercorrelations for the 27 questions. The three factors which accounted for the intercorrelations were named: reasoned determination to achieve the objective, confidence in leaders and satisfaction with traditional values. The importance of these factors can be judged by the average variance thus accounted for: 11%, 8%, and 6% respectively. The average

uniqueness variance is 75% of the total. In reality, this means that the battery contains three group or common factors of minor importance and 27 specific factors. For only one question is the communality greater than the uniqueness. For half the questions, the communality is less than .25, which is what one would expect when the average of the absolute values of the intercorrelations is only .14. This leads one to suspect the meaning of "generalized" attitudes as discussed on page 272. It should be borne in mind, however, that one reason for the lack of any marked general factor or factors in such a battery is the undoubtedly low reliabilities for the responses being correlated.

The most recent article on morale measurement is that of Estes and Estes (**34**) who propose "miniature" scales, of from 3 to 8 questions each, for measuring 11 aspects of morale. The test-retest reliabilities range from .51 to .91, median of .82, hence it is doubtful whether the "scales are sufficiently reliable for use in tracing changes in attitudes." One certainly would be skeptical of individual changes. The total score based on all 47 items yielded a reliability of .93, but the meaning of total scores is obscured by the very low interrelationships among the 11 aspects—for two samples, only 6 of the 55 intercorrelations are consistently greater than .30. The authors state that all the aggressive attitudes tend to go together, but examination of their data indicates that this holds for only one of their samples.

What conclusions can be drawn from this brief review of the attempts to measure and analyze morale? It seems obvious that morale is not an entity, that there are many "morales," that there has been an open season for calling a wide variety of things by the name of morale, that the several aspects of morale can be fairly reliably measured by scale techniques, that nothing is known about the reliability of single question indicators of morale, that the instruments for gauging morale pertaining to war effort are soon outdated, that some types of morale are mutually correlated though in varying degrees, that little is to be gained by factor analyzing intercorrelations based on single and likely unreliable questions, and that much of the available research on morale suffers from a lack of hard-boiled critical mindedness.

It may be that "morale is a lot of little things," hence difficult to define as a scientific concept. It seems more reasonable, however, to believe that these "little things" are not entirely independent, either statistically or functionally, that certain of them tend to go together or form clusters, and that such clusters are conceivably independent of each other. Validation of scales designed to measure these several factors would require as many external criteria as there are factors. Two jobs need to be done: *the determination of the dimensions of morale and the construction and validation of scales for measuring these dimensions.*

Where to begin is a difficult question, but the following procedure is suggested. Start with postulated components. Here we need to make a distinction between two concepts. Although "component" and "factor" are frequently used synonymously, we shall for the purpose of this discussion use the term "factor" as in current factor analysis, i.e., "factors" are statistically independent in the sense that they are the orthogonal axes or dimensions necessary to explain a set of intercorrelations. If certain oblique axes are needed, then certain factors would not be uncorrelated. By "components" we shall mean those aspects which are logically postulated as the ingredients of morale. It is not implied that these postulated components are mutually independent. They constitute the many and varied "morales" which have already been defined in the literature and any additional concepts which workers in this area care to propose. Each postulated component should be carefully defined. Perhaps it would be necessary to call in an expert in semantics in order to determine whether certain definitions were really different. There is no way of knowing how many components might be postulated. If an unwieldly number, say more than 60, were proposed, it would be advisable to submit the list to a group of competent judges to see if some might be eliminated. It is certainly not unreasonable to require that a proposal should have objective acceptance beyond the subjective horizon of its proponent.

The next step would be to construct a uni-dimensional scale for measuring each of the several components. Use of the Guttman (52) technique would greatly facilitate the meeting of this requirement. Scales are suggested rather than single questions so as to have reliable measures. A Guttman type scale for each component would also avoid the absurdity of having several questions which supposedly tap a given component turn out, as sometimes happened in the Cantril study (18), to have little or nothing in common.

When reliable uni-dimensional scales have been constructed, they should be administered to a cross section of the adult population. A carefully stratified sample of 400 cases would be sufficient for determining the intercorrelations of the several scales. It would be more economical to compute the correlations by the product moment method than to have to double the size of the sample in order to have the same degree of stability for tetrachoric coefficients. The table of correlations among the postulated components would then be factor analyzed to ascertain how many factors underlie the several aspects of morale. If the centroid method were used, rotations would be necessary and there would be some difficulty in finding appropriate descriptive labels for the factors. Then scales could be constructed for measuring each of these "primary"

factors, validity studies could be made, and finally a search for psychosocial correlates of the "morales" could be undertaken.

This may sound like an ambitious program but it would be one way to bring some order into a rather chaotic field of research.

IX. Concluding Remarks

An attempt has been made herein to provide a critical appraisal of what the writer considers to be the chief methodological problems involved in research on attitudes and opinions. It has been assumed that progress in developing a science of attitudes and opinions, or the fruitful scientific study thereof as parts of whatever discipline, will continue to be largely dependent upon the success with which attitude and opinion variables are measured. In this connection, we have stressed the basic need for reliability, validity, and uni-dimensionality for the instruments or devices used to classify or measure individuals with respect to their opinions or attitudes. Attitude scales can be so constructed as to attain satisfactory reliability. Unitary scales can be developed by the Guttman scaling technique. Validity can be established, but not without more effort than is usually expended.

Opinion gaugers have paid scant attention to these three basic requirements. They have, however, demonstrated that the general dependability of the single question method is none too high—a variety of percentage variations can be produced by changing the wording and form of the question and by altering the response setup. Enough is now known concerning the possible variations associated with question-answer form to justify the conclusion that the typical poll results are subject to a variety of possible errors. Because of its many grave difficulties, we have suggested that the single question technique be replaced by attitude scales. Has a science ever suffered because of refinements in measuring techniques? It is true, of course, that less research would be possible if scales were substituted for single questions, but one good study might be worth several based on inadequate techniques.

The statistical issues in attitude-opinion research are not different from those encountered in other fields of social science. Inadequate analyses and statistical errors have been plentiful, but as more statistical sophistication is acquired, one can expect adequate statistical treatment with fewer errors. Securing representative samples from the general population is no longer an impossible task, but the study of attitudes has been so limited by the campus-bound inertia of researchers that generalizations have tended to hold only for college students, rather than for man in general.

A study of changes under experimental conditions with necessary control groups would seem to be a fruitful way of ascertaining some of the factors which contribute to the formation or learning of attitudes and opinions. Lifelike experiences can be provided in order to see whether changes take place. The other principal method for finding attitude and opinion determiners is to look for correlates. This search could be facilitated by a greater stress on the critical formulation of hypotheses.

We have suggested that the judicious use of factor analysis might help to bring some order out of the chaos in certain areas, particularly that of morale and of liberalism-conservatism.

This appraisal has not touched on certain fields of application of attitude-opinion techniques, e.g., market and consumer research, which the reader will find discussed in Blankenship's *Consumer and Opinion Research* (**10**). Nor have we been concerned with all the problems associated with election polling. The reader is referred to Gallup and Rae's *Pulse of Democracy* (**49**) for greater detail on this topic. The serious student will not be aided by Gallup's recent publication, *A Guide to Public Opinion Polls* (**48**), whereas he will definitely benefit from a study of the analyses by Katz (**64, 66**) of the 1940 and 1944 polls.

We have not found an integrated summary of the findings of public opinion research. Newcomb (**93**) has provided an excellent summary and synthesis of attitude studies up to 1937, with a good discussion of methodological issues.

It will have been noted that very little has been said herein concerning the work on opinions and attitudes by the various governmental agencies, such as Program Surveys (under Rensis Likert) in the Bureau of Agricultural Economics, OWI's Surveys Division (Keith Kane, Rensis Likert, Julian Woodward, Elmo Wilson, Daniel Katz, NORC, *et al.*), and the Research Branch of the Army's Morale Services Division (Samuel Stouffer, Leonard Cottrell, Carl Hovland, *et al.*). Certain individuals in these agencies have given such loud acclaim to their interest in and work on methodological problems that some social scientists have been quite optimistic about the contributions to techniques which might emerge from this extensive work. Can it be said that these agencies have developed and perfected techniques which will aid in the solution of problems in the attitude-opinion field?

The writer had close contact, from 1941 through the winter of 1944, as a consultant to one of these agencies (the Army's) and in order to have a broader basis for answering the above question he has conferred with key men in the other agencies. All were very cooperative and willingly talked over their methodological problems, but a documented

See pg 570

answer to our question is impossible because so little concerning the work on methodological issues can be found in the records of the agencies —the available information seems to be cached in the minds of staff members. It is apparent, therefore, that any appraisal must be tentative; yet the writer feels reasonably sure that the following observation is irrefutable: no marked contributions to methodology were made during the first 30 months of the war effort.

Minor advances have been made. The Census Bureau and Likert's group in the Department of Agriculture have, as previously noted, succeeded in ironing out the biasing bugs of sampling, but this was being done before the United States entered the war. Likert has made progress in refining open-end question interviewing. Guttman's scaling method originated in connection with his work as a consultant to the Research Branch of the Army's Morale Services Division. When completed, this method will be superior to the equal appearing intervals and internal consistency techniques, and will be applicable to other types of psychological measurement. All the major agencies have carried out studies of a methodological nature, but the striking thing is the fewness and limited nature of such studies. One looks in vain for work on the fundamental problems of reliability and validity. No doubt many individuals in these agencies have acquired a wealth of information and experience which may later lead to major improvements along methodological lines.

The above remarks are not to be interpreted as an indictment—perhaps those who profess their interest in methodology have been unable to do much because of time pressure. Even those who have hoped that their work will result in residual contributions to scientific techniques have found that their active interest in studying methods tends to wax and wane. Temporary spurts of enthusiasm were often dwarfed by the immediate problems of the day, and sometimes dampened by budgetary difficulties. A reading of the recent paper by Woodward (131) makes it quite evident that those in government who were doing opinion research during the war were definitely handicapped in various ways. The writer hopes that he is in error in failing to see any reason for optimism concerning war-born advances in the attitude-opinion field of research.

See pg 570 #62

BIBLIOGRAPHY

1. ALLPORT F. H., & HARTMAN, D. A. The measurement of atypical opinion in a certain group. *Amer. polit. Sci. Rev.*, 1925, **19**, 735–760.
2. BALLIN, M. R., & FARNSWORTH, P. R. A graphic rating method for determining the scale values of statements in measuring social attitudes. *J. soc. Psychol.*, 1941, **13**, 323–327.

3. BENSON, L. E. Studies in secret-ballot technique. *Publ. Opin. Quart.*, 1941, **5**, 79–82.
4. BEYLE, H. C. A scale for the measurement of attitude toward candidates for elective governmental office. *Amer. polit. Sci. Rev.*, 1932, **26**, 527–544.
5. BLANKENSHIP, A. B. Methods of measuring industrial morale. In G. W. Hartmann & T. Newcomb (Eds.), *Industrial conflict; a psychological interpretation*. New York: Cordon, 1939. Pp. 299–312.
6. BLANKENSHIP, A. B. The choice of words in poll questions. *Sociol. soc. Res.*, 1940, **25**, 12–18.
7. BLANKENSHIP, A. B. Does the question form influence public opinion poll results? *J. appl. Psychol.*, 1940, **24**, 27–30.
8. BLANKENSHIP, A. B. The effect of the interviewer upon the response in a public opinion poll. *J. consult. Psychol.*, 1940, **4**, 134–136.
9. BLANKENSHIP, A. B. The influence of the question form upon the response in a public opinion poll. *Psychol. Rec.*, 1940, **3**, 345–422.
10. BLANKENSHIP, A. B. *Consumer and opinion research*. New York: Harper, 1943.
11. BRUNER, J. S. How much post-war migration? *Amer. J. Sociol.*, 1943, **49**, 39–45.
12. BRUNER, J. S. Public thinking on post-war problems. Washington: Nat'l Planning Ass., Planning Pamphlet No. 23, 1943. Pp. 36.
13. BUGELSKI, R., & LESTER, O. P. Changes in attitudes in a group of college students during their college course and after graduation. *J. soc. Psychol.*, 1940, **12**, 319–332.
14. CAHALAN, D., & MEIER, N. C. The validity of mail ballot polls. *Psychol. Rec.*, 1939, **3**, 2–11.
15. CANTRIL, H. Experiments in the wording of questions. *Publ. Opin. Quart.*, 1940, **4**, 330–332.
16. CANTRIL, H. America faces the war: a study in public opinion. *Publ. Opin. Quart.*, 1940, **4**, 387–407.
17. CANTRIL, H. Public opinion in flux. *Ann. Amer. Acad. polit. soc. Sci.*, 1942, **220**, 136–152.
18. CANTRIL, H. *Gauging public opinion*. Princeton: Princeton Univ. Press, 1944.
19. CANTRIL, H. Do different polls get the same results? *Publ. Opin. Quart.*, 1945, **9**, 61–69.
20. CANTRIL, H., & HARDING, J. The 1942 elections: a case study in political psychology. *Publ. Opin. Quart.*, 1943, **7**, 222–241.
21. CANTRIL, H., & RUGG, D. Looking forward to peace. *Publ. Opin. Quart.*, 1940, **4**, 119–121.
22. CHANT, S. N. F., & MYERS, C. R. An approach to the measurement of mental health. *Amer. J. Orthopsychiat.*, 1936, **6**, 134–140.
23. CHILD, I. L. Morale: a bibliographical review. *Psychol. Bull.*, 1941, **38**, 393–420.
24. CONNELLY, G. M. Now let's look at the real problem: validity. *Publ. Opin. Quart.*, 1945, **9**, 51–60.
25. CONRAD, H. S., & SANFORD, R. N. Scales for the measurement of war-optimism: I. Military optimism; II. Optimism on consequences of the war. *J. Psychol.*, 1943, **16**, 285–311.
26. CONRAD, H. S., & SANFORD, R. N. Some specific war-attitudes of college students. *J. Psychol.*, 1944, **17**, 153–186.
27. COREY, S. M. Professed attitudes and actual behavior. *J. educ. Psychol.*, 1937, **28**, 271–280.
28. COREY, S. M. Changes in the opinions of female students after one year at a university. *J. soc. Psychol.*, 1940, **11**, 341–351.
29. CRONBACH, L. J. *Exploring the wartime morale of high-school youth*. Stanford Univ.: Stanford Univ. Press, 1943.

30. DARLEY, J. G. Changes in measured attitudes and adjustments. *J. soc. Psychol.*, 1938, **9**, 189–199.

31. DIGGS, E., HANGER, E., & MULL, H. K. Morale in the college situation in relation to the morale scale of Rundquist and Sletto. *Amer. J. Psychol.*, 1942, **55**, 561–562.

32. DUDYCHA, G. J. A critical examination of the measurement of attitude toward war. *J. soc. Psychol.*, 1943, **18**, 383–392.

33. ERICKSEN, S. C. A skeptical note on the use of attitude scales toward war: I. In 1940, 1941. *J. soc. Psychol.*, 1942, **16**, 229–242.

34. ESTES, W. K., & ESTES, K. W. Minnesota studies in war psychology: I. A set of miniature scales for the measurement of attitudes related to morale. *J. soc. Psychol.*, 1944, **20**, 265–276.

35. EWING, T. N. A study of certain factors involved in changes of opinion. *J. soc. Psychol.*, 1942, **16**, 63–88.

36. FARNSWORTH, P. R. Changes in "attitude toward war" during the college years. *J. soc. Psychol.*, 1937, **8**, 274–279.

37. FARNSWORTH, P. R. Shifts in the values of opinion items. *J. Psychol.*, 1943, **16**, 125–128.

38. FERGUSON, L. W. An item analysis of Peterson's "war" scale. *Psychol. Bull.*, 1938, **35**, 521.

39. FERGUSON, L. W. Primary social attitudes. *J. Psychol.*, 1939, **8**, 217–223.

40. FERGUSON, L. W. The measurement of primary social attitudes. *J. Psychol.*, 1940, **10**, 199–205.

41. FERGUSON, L. W. The stability of the primary social attitudes: I. Religionism and humanitarianism. *J. Psychol.*, 1941, **12**, 283–288.

42. FERGUSON, L. W. A study of the Likert technique of attitude scale construction. *J. soc. Psychol.*, 1941, **13**, 51–57.

43. FERGUSON, L. W. The isolation and measurement of nationalism. *J. soc. Psychol.*, 1942, **16**, 215–228.

44. FERGUSON, L. W. The relation of the primary social attitude variables to "national morale." *Amer. sociol. Rev.*, 1944, **9**, 194.

45. FERGUSON, L. W. A revision of the primary social attitude scale. *J. Psychol.*, 1944, **17**, 229–241.

46. FERGUSON, L. W. Socio-psychological correlates of the primary attitude scales: I. Religionism; II. Humanitarianism. *J. soc. Psychol.*, 1944, **19**, 81–98.

47. FIELD, H. H., & CONNELLY, G. M. Testing polls in official election booths. *Publ. Opin. Quart.*, 1942, **6**, 610–616.

48. GALLUP, G. *A guide to public opinion polls.* Princeton: Princeton Univ. Press, 1944.

49. GALLUP, G., & RAE, S. F. *The pulse of democracy.* New York: Simon & Schuster, 1940.

50. GHISELLI, E. E. All or none *versus* graded response questionnaires. *J. appl. Psychol.*, 1939, **23**, 405–413.

51. GOSNELL, H. F. How accurate were the polls? *Publ. Opin. Quart.*, 1937, **1**, No. 1, 97–104.

52. GUTTMAN, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, **9**, 139–150.

53. HALL, O. M. Attitudes and unemployment. A comparison of the opinions and attitudes of employed and unemployed men. *Arch. Psychol., N. Y.*, 1934, **25**, No. 165, 1–65.

54. HANSEN, M. H., & HAUSER, P. M. Area sampling—some principles of sample design. *Publ. Opin. Quart.*, 1945, **9**, 183–193.

55. HARDING, J. A scale for measuring civilian morale. *J. Psychol.*, 1941, **12**, 101–110.

56. HAYES, S. P., JR. The inter-relations of political attitudes: II. Consistency in voters' attitudes. III. General factors in political issues. IV.

Political attitudes and party regularity. *J. soc. Psychol.*, 1939, **10**, 359–398; 503–552.

57. HILGARD, E. R., & PAYNE, S. L. Those not at home: riddle for pollsters. *Publ. Opin. Quart.*, 1944, **8**, 254–261.
58. HOROWITZ, E. L. The development of attitude toward the Negro. *Arch. Psychol., N. Y.*, 1936, **28**, No. 194, 1–47.
59. HULL, R. L., & KOLSTAD, A. Morale on the job. In G. Watson (Ed.), *Civilian Morale.* Boston: Houghton Mifflin, 1942. Pp. 349–364.
60. HUNTER, E. C. Changes in general attitudes of women students during four years in college. *J. soc. Psychol.*, 1942, **16**, 243–257.
61. HYMAN, H. Do they tell the truth? *Publ. Opin. Quart.*, 1944, **8**, 557–559.
62. JOHNSON, H. M. Pseudo-mathematics in the mental and social sciences. *Amer. J. Psychol.*, 1936, **48**, 342–351.
63. JONES, V. Attitudes of college students and the changes in such attitudes during four years in college. *J. educ. Psychol.*, 1938, **29**, 14–25; 114–134.
64. KATZ, D. The public opinion polls and the 1940 election. *Publ. Opin. Quart.*, 1941, **5**, 52–78.
65. KATZ, D. Do interviewers bias poll results? *Publ. Opin. Quart.*, 1942, **6**, 248–268.
66. KATZ, D. The polls and the 1944 election. *Publ. Opin. Quart.*, 1944, **8**, 468–482.
67. KIRKPATRICK, C. Assumptions and methods in attitude measurement. *Amer. sociol. Rev.*, 1936, **1**, 75–88.
68. KIRKPATRICK, C., & STONE, S. Attitude measurement and the comparison of generations. *J. appl. Psychol.*, 1935, **19**, 564–582.
69. KOLSTAD, A. Employee attitudes in a department store. *J. appl. Psychol.*, 1938, **22**, 470–479.
70. LA PIERE, R. T. Attitudes vs. actions. *Social Forces*, 1934, **13**, 230–237.
71. LAZARSFELD, P. F. The change of opinion during a political discussion. *J. appl. Psychol.*, 1939, **23**, 131–147.
72. LAZARSFELD, P. F. Panel studies. *Publ. Opin. Quart.*, 1940, **4**, 122–128.
73. LAZARSFELD, P. F. The controversy over detailed interviews—an offer for negotiation. *Publ. Opin. Quart.*, 1944, **8**, 38–60.
74. LAZARSFELD, P., & FISKE, M. The "panel" as a new tool for measuring opinion. *Publ. Opin. Quart.*, 1938, **2**, 596–612.
75. LAZARSFELD, P. F., BERELSON, B., & GAUDET, H. *The people's choice.* New York: Duell, Sloan, and Pearce, 1944.
76. LENTZ, T. F. Personage admiration and other correlates of conservatism-radicalism. *J. soc. Psychol.*, 1939, **10**, 81–93.
77. LIKERT R. A technique for the measurement of attitudes. *Arch. Psychol., N. Y.*, 1932, **22**, No. 140, 1–55.
78. LIKERT, R., ROSLOW, S., & MURPHY, G. A simple and reliable method of scoring the Thurstone attitude scales. *J. soc. Psychol.*, 1934, **5**, 228–238.
79. LINK, H. C. The eighth nation-wide social experimental survey. *J. appl. Psychol.*, 1943, **27**, 1–11.
80. LINK, H. C. An experiment in depth interviewing on the issue of internationalism vs. isolationism. *Publ. Opin. Quart.*, 1943, **7**, 267–279.
81. LORGE, I. The Thurstone attitude scales: I. Reliability and consistency of rejection and acceptance. *J. soc. Psychol.*, 1939, **10**, 187–198.
82. MATHEWS, C. O. The effect of printed response words upon children's answers to questions in two-

response types of tests. *J. educ. Psychol.*, 1927, **18**, 445–457.

83. MAY, M. A., & HARTSHORNE, H. First steps toward a scale for measuring attitudes. *J. educ. Psychol.* 1926, **17**, 145–162.
84. MCNEMAR, Q. A critical examination of the University of Iowa studies of environmental influences upon the IQ. *Psychol. Bull.*, 1940, **37**, 63–92.
85. MCNEMAR, Q. Sampling in psychological research. *Psychol. Bull.*, 1940, **37**, 331–365.
86. MERTON, R. K. Fact and factitiousness in ethnic opinionnaires. *Amer. sociol. Rev.*, 1940, **5**, 13–28.
87. MILLER, D. C. Morale of college-trained adults. *Amer. sociol. Rev.*, 1940, **5**, 880–889.
88. MILLER, D. C. Personality factors in the morale of college-trained adults. *Sociometry*, 1940, **3**, 367–382.
89. MILLER, D. C. Economic factors in the morale of college-trained adults. *Amer. J. Sociol.*, 1941, **47**, 139–156.
90. MILLER, D. C. The measurement of national morale. *Amer. sociol. Rev.*, 1941, **6**, 487–498.
91. MILLER, D. C. National morale of American college students in 1941. *Amer. sociol. Rev.*, 1942, **7**, 194–213.
92. MILLER, D. C. Effect of the war declaration on the national morale of American college students. *Amer. sociol. Rev.*, 1942, **7**, 631–644.
93. MURPHY, G., MURPHY, L. B., & NEWCOMB, T. M. *Experimental social psychology*. New York: Harper, 1937.
94. NEELY, T. E. *A study of error in the interview*. (Privately publ.), 1937. Pp. 150.
95. NEWCOMB, T. Labor unions as seen by their members: an attempt to measure attitudes. In G. W. Hartmann & T. Newcomb (Eds.), *Industrial conflict: a psychological interpretation*. New York: Cordon, 1939. Pp. 313–348.
96. NEWCOMB, T. M. *Personality and social change*. New York: Dryden Press, 1943.
97. PACE, C. R. A situations test to measure social-political-economic attitudes. *J. soc. Psychol.*, 1939, **10**, 331–344.
98. PROSHANSKY, H. M. A projective method for the study of attitudes. *J. abnorm. soc. Psychol.*, 1943, **38**, 393–395.
99. REMMERS, H. H., *et al.* Studies in attitudes. *Purdue Univ. Stud. higher Educ.*, 1934, **25**, 1–112.
100. REMMERS, H. H., *et al.* Further studies in attitudes, Series II. *Purdue Univ. Stud. higher Educ.*, 1936, **31**, 1–298.
101. REMMERS, H. H., *et al.* Further studies in attitudes, Series III. *Purdue Univ. Stud. higher Educ.*, 1938, **34**, 1–151.
102. REMMERS, H. H., & SILANCE, E. B. Generalized attitude scales. *J. soc. Psychol.*, 1934, **5**, 298–312.
103. REUSS, C. F. Differences between persons responding and not responding to a mailed questionnaire. *Amer. sociol. Rev.*, 1943, **8**, 433–438.
104. RIKER, B. L. A comparison of methods used in attitude research. *J. abnorm. soc. Psychol.*, 1944, **39**, 24–42.
105. RIKER, B. L. Comparison of attitude scales—a correction, *J. abnorm. soc. Psychol.*, 1945, **40**, 102–103.
106. ROPER, E. Wording of questions for the polls. *Publ. Opin. Quart.*, 1940, **4**, 129–130.
107. ROPER, E. Checks to increase polling accuracy. *Publ. Opin. Quart.*, 1941, **5**, 87–90.
108. ROSANDER, A. C. An attitude scale based upon behavior situations. *J. soc. Psychol.*, 1937, **8**, 3–15.
109. ROSLOW, S., WULFECK, H., & CORBY, G. Consumer and opinion research: experimental studies on the form of the question. *J. appl. Psychol.*, 1940, **24**, 334–346.

110. RUGG, D. Experiments in wording questions: II. *Publ. Opin. Quart.*, 1941, **5**, 91–92.
111. RUGG, D., & CANTRIL, H. The wording of questions in public opinion polls. *J. abnorm. soc. Psychol.*, 1942, **37**, 469–495.
112. RUNDQUIST, E. A., & SLETTO, R. F. *Personality in the depression.* Minneapolis: Univ. Minn. Press, 1936.
113. SANFORD, R. N., & CONRAD, H. S. Some personality correlates of morale. *J. abnorm. soc. Psychol.*, 1943, **38**, 3–20.
114. SCHANCK, R. L. A study of a community and its groups and institutions conceived of as behaviors of individuals. *Psychol. Monogr.*, 1932, **43**, No. 2, 1–133.
115. SEASHORE, R. H., & HEVNER, K. A time-saving device for the construction of attitude scales. *J. soc. Psychol.*, 1933, **4**, 366–372.
116. SHUTTLEWORTH, F. K. Sampling errors involved in incomplete returns to mail questionnaires. *J. appl. Psychol.*, 1941, **25**, 588–591.
117. SKOTT, H. E. Attitude research in the Department of Agriculture. *Publ. Opin. Quart.*, 1943, **7**, 280–292.
118. STAGNER, R. Fascist attitudes: an exploratory study. *J. soc. Psychol.*, 1936, **7**, 309–319.
119. STANTON, F. Notes on the validity of mail questionnaire returns. *J. appl. Psychol.*, 1939, **23**, 95–104.
120. STONBOROUGH, T. H. W. The continuous consumer panel: a new sampling device in consumer research. *Appl. Anthrop.*, 1942, **1**, No. 2, 37–41.
121. SUCHMAN, E. A., & MCCANDLESS, B. Who answers questionnaires? *J. appl. Psychol.*, 1940, **24**, 758–769.
122. THURSTONE, L. L. Attitudes can be measured. *Amer. J. Sociol.*, 1928, **33**, 529–554.
123. THURSTONE, L. L., & CHAVE, E. J. *The measurement of attitude.* Chicago: Univ. Chicago Press, 1929.
124. UDOW, A. B. The "interview-effect" in public opinion and market research surveys. *Arch. Psychol., N. Y.*, 1942, No. 277, 1–36.
125. WANG, K. A. Suggested criteria for writing attitude statements. *J. soc. Psychol.*, 1932, **3**, 367–373.
126. WATSON, G. (Ed.) *Civilian morale.* Boston: Houghton Mifflin, 1942.
127. WATSON, G. Morale during unemployment. In G. Watson (Ed.), *Civilian Morale.* Boston: Houghton Mifflin, 1942. Pp. 273–348.
128. WHISLER, L. D., & REMMERS, H. H. Liberalism, optimism, and group morale: a study of student attitudes. *J. soc. Psychol.*, 1938, **9**, 451–467.
129. WILKS, S. S. Representative sampling and poll reliability. *Publ. Opin. Quart.*, 1940, **4**, 261–269.
130. WILKS, S. S. Confidence limits and critical differences between percentages. *Publ. Opin. Quart.*, 1940, **4**, 332–338.
131. WOODWARD, J. L. Making government opinion research bear upon operation. *Amer. sociol. Rev.*, 1944, **9**, 670–677.
132. YOUNG, K. (Ed.) *Social attitudes.* New York: Holt, 1931.
133. ZUBIN, J., & GRISTLE, M. An empirical scale for measuring militarism-pacificism. *Psychol. Rec.*, 1937, **1**, 27–32.

REPLY TO ELDER'S NOTE ON DUNLAP'S REMEDY FOR COLOR VISION

KNIGHT DUNLAP
University of California (Los Angeles)

In the note by Elder,* I found one item with which I agree completely, namely "The difficulties of conducting research from an academic base on a shifting war-time population." The difficulty was extreme, but we continued our research as a contribution to the armed services and the men therein.

As to our use of the American Optical Company's chart-test in spite of its unreliability, we used it because it was the test the men were required to pass for acceptance in certain of the services.

Concerning the Louisiana State University experiment, I did not reject the results because of unknown quantities of vitamin A, but because, as indicated in Elder's preliminary report (not included in the bibliography of the *Note*) the material was not pure vitamin A, but provitamin A.

We have long known that a "color-blind" person is unable to transform that provitamin into the vitamin, and we have had to abandon our research on so-called "color-blindness" because of the impossibility of obtaining vitamin A. Some pharmaceutical firms label their product Vitamin A, in large print, and in smaller print admit that it is a mixture of vitamin A, without indicating the proportions of the two factors. That the products of other firms, labeled Vitamin A, are mixtures of the same sort, is quite probable, even if it is not admitted.

As to Elder's suggestions for my future work: (a) was fulfilled in the work reported; (b) supervised administration was difficult for men in the services; (c) complete treatment was applied to the point at which some passed the service test; and (d) there was no possibility of the persons learning the tests, because they were not told whether their readings were right or wrong, or whether their judgments on the Nela test were right or wrong. As to Elder's statement that "on the basis of results presented on tests before and after treatment, he should have rejected 14 out of 14," I am not in agreement. As to the preceding statement: "If the summary statements given on each case may be interpreted literally then Dunlap himself rejects 12 of the 14 cases as questionable." I still disagree with Elder. Improvement was shown in almost all cases, and since the purpose of the treatment was to enable the individuals to pass the service tests, we could not go beyond that point. This reenforces my opening quotation from Elder's *Note*. The difficulty was indeed great.

* Elder, J. H. Note on Dunlap's remedy for color vision. *Psychol. Bull.*, 1946. **43**, 77–79.

BOOK REVIEWS

WERTHEIMER, MAX *Productive thinking*. New York: Harper & Bros., 1945. Pp. xiv+224.

America often reaps a harvest in providing haven for refugees. This time America—and the world—is richer for Max Wertheimer's studies carried out at the New School for Social Research. His book consists largely of concrete illustrations of "genuine, fine, clean, direct productive processes" given in great detail. In chapters I through III Wertheimer treats three relatively simple mathematical problems: finding the area of a parallelogram, proving the equality of the vertical angles formed by two intersecting straight lines, and Gauss' theorem. In chapter V he recounts the development of his own discoveries of the sum of the angles of a polygon, both plane and solid. Chapters VI and VII are psychological reconstructions of Galileo's discovery of the law of inertia and Einstein's thinking that led to the theory of relativity. In the concluding chapter Wertheimer summarizes the dynamic nature of the creative process, which he had slowly elaborated as he treated the various examples, as follows:

> When one grasps a problem situation, its structural features and requirements set up certain strains, stresses, and tensions in the thinker. What happens in real thinking is that these strains and stresses are followed up, yield vectors in the direction of improvement of the situation, and change it accordingly.

The problems thus far mentioned, with the exception of Einstein's, are relatively self-contained, closed systems. In chapter IV Wertheimer considers two concrete social situations in which the field is no longer limited to the internal forces in the objective situation, but includes concrete ego-tendencies and personal needs. In such cases, the solution arises out of transformations of the relation between the problem-situation and the ego. In each transformation, he claims the "structural features remain decisive."

On occasion, after posing a question suggested by the concrete problem under consideration, he answers in terms of logic and association theory. Then he demonstrates their inadequacy. Immediately he follows with an explanation based upon the structural characteristics of the situation and shows how the central core of the question now first becomes answered. Although the traditional theories at times are "adequate and useful," Wertheimer believes they are generally blind to the dynamics in thinking. They "show a dead picture stripped of all that is alive in them."

This book is offered to the psychologist, logician, and educator "as an invitation to reconsider basic issues." It is loaded with provocation. The educator is challenged to teach so that "genuine discovery in tasks

of a structural nature plays an essential role." Emphasis on mechanical drill is "dangerous because it easily induces habits of sheer mechanized action, blindness, tendencies to perform slavishly instead of thinking, instead of facing a problem freely." The logician too is often taken to task for his structure-blindness.

There is a sparkle and warmth about this book that is captivating. Rarely does the psychological writer actually use the principles he extolls in the very construction of his book. Wertheimer does. For instance, he invites penetration, "Think it over yourself reader . . . Try to do justice to the points I am going to mention." The diagrams and illustrations themselves are used with remarkable effectiveness. The printer has placed them so expertly in the text that few formal references to them are required. They integrate smoothly into the flow of thought in the passage. Perhaps the quality which preserves the book's freshness from beginning to end is Wertheimer's attitude that he has highly satisfying and enjoyable insights to share with the reader. The book is gaily decked with such sentences as, "Try to do it before reading what I report in this chapter. You may enjoy what follows more."

Yet, when you have finished the book, you have no "clear" insight, you know that the dynamics of productive thinking are still far from being structurally "clean." No attempt is made to define concisely such terms as "inner structure." Wertheimer recognized and humbly admits this: "My terms should not give the impression that the problems are settled; they themselves are loaded with—I think—productive problems."

Scattered throughout the book are many suggestions for such research work. Wertheimer himself refers to a few odds and ends of systematic research he carried out. He does not give his methods or results explicitly, however, as he centers his full attention throughout the book upon the qualitative aspects of the productive process itself. It is curious that he never once refers to Duncker's *Zur Psychologie des productiven Denkens* (1935), which bears such close resemblance to his book.

The main value of Wertheimer's book will rest in its live challenge to psychologists to work in the field of creative thought, stressing the dynamic aspects. He wants contemporary psychologists to "look at the situation freely, open-mindedly, . . . to penetrate, to realize and trace out the inner" relationships of the thought process. The crucial test of the power of Wertheimer's invitation will be found in the number and quality of research workers who accept his challenge. Let us hope there will be many, for our world needs "clean" thinking. Wertheimer has done a remarkable job of blazing the trail; the way is no longer uncharted.

HAROLD GUETZKOW.

University of Michigan.

LECKY, PRESCOTT. *Self-consistency, a theory of personality.* New York: Island Press, 1945. Pp. 154.

This posthumous collection of essays, edited by John Taylor and supplied with an appreciative foreword of the author and his special psychological contributions by Gardner Murphy, develop the radical proposition that the individual "defines" for himself the nature of the totality which he is. New experiences are assimilated only when they are viewed as forward steps continuing and fulfilling oneself. This basic thesis rejects all the customary versions of the learning process, makes the organism its own determiner, and accepts but a single "general factor," viz., the personality itself.

Inventory data derived from Columbia College students indicate that confidence in social situations, optimism, a friendly attitude toward other members of the family, companionship with the opposite sex, efficient work dispositions, freedom from nervous symptoms, and a feeling of physical well-being are closely interrelated and tend to increase and decrease together. This is construed as fatal to the specificity doctrine. Instead, predictability of behavior is a function of stability which arises out of the basic need or general motive for consistent self-organization—the organism's sole purpose.

This position is supported by interpreting *learning* as always a resolution of conflict (reduction to unity), *pleasure* as difficult unification of a manifold (the barely successful synthesis), *emotion* as evidence of the effort required to achieve or preserve unity, and *personality* as an organization of values consistent with each other. Freud is critized for not recognizing that the patient's resistance is prerequisite to his integrity; *therapy must shift from the effort to make the patient consistent with society to making him consistent with himself.* Educational and allied defects are viewed as consequences of inhibiting "definitions" of our own natures from which we can be freed as soon as we see them as burdens and not as assets.

Lecky's clinical practice was apparently quite successful in using this clear-cut approach with its direct appeal to the client's standards of "belongingness." He asks psychologists to weigh the implications of their own "system-building" as evidence for his position. "The problems of science are not abstractions, but the human problems of scientists. If we did not strive for consistency in ourselves, what possibly could be the motive for seeking and demanding consistency in abstract scientific formulations?" (112).

These meaty papers are rich in both critical ideas and creative concepts. Lecky's abbreviated analysis resembles Allport's extended treatment at a number of points minus his trait theory and particularly the "eclectisicm" which is prima facie "inconsistent." Its positive features are an appealing simplicity, optimism about the possibilities of re-

directing human energies, and ready availability for the tasks of the practitioner.

Nonetheless, grave skepticism is warranted at many points. Is the delinquent primarily such because he conceives of himself as a "bad boy"? Does one become honest (or a charming wife, or a keen lawyer, etc.) merely by affirming that such shall hereafter be part of one's nuclear self? Functional inadequacies here are purely peripheral; but what if the "central core" is diseased? The claim that one is or becomes largely what one "pictures" onself to be does not seem to square with the grim facts of frustration. Had Lecky lived to complete this theoretical structure, these questions might have received appropriate consideration. As it is, we must be grateful to him for his effective insistence that all individual behavior is a story with a plot wherein the evolving value-pattern of the actor is the most potent single psychological reality.

GEORGE W. HARTMANN.

Teachers College, Columbia University.

BECK, S. J. *Rorschach's Test: 1. Basic processes.* New York: Grune & Stratton, 1944. Pp. xiii+223.

It is unquestionably true that Dr. Beck has once more demonstrated his competence and scholarship by producing this book. The question, however, is just how much is contributed by it—and by other like material—to American psychology and psychiatry. From where this reviewer sits, the answer to the question is disappointing.

Volume 1 of Beck's *Rorschach's Test* proposes to bring the presumptive reader up to date in Dr. Beck's cherished ambition to standardize and "demonstrate the processes used in evaluating Rorschach test responses." It fulfills its announced intention by painstakingly and deliberately detailing responses, rationalizing the use of the various symbols, explaining and cataloguing, arguing and justifying, through sixteen laborious chapters chaperoned by a brief bibliography, two appended tables, and ten disjointed schema of the blots. With commentary added to commentary, as one reads it, it takes on more and more the character of the Babylonian *Talmud*. To read such a book is obviously out of the question: to study it requires a fixity of purpose amounting almost to a dedication. Its real service will be as another standard reference work in a young but already needlessly over-complicated field.

The practicing psychiatrist or psychologist is coming to rely upon the Rorschach as he does upon almost no other tool. It is a good technique, perhaps the best of the projective methods, and he appreciates what it can do for him by way of evaluating the personality and, indeed, for certain services nowhere else available. He regards it

practically, as an invaluable instrument for diagnosis, as an index and guide to therapeusis, as a clue to treatment progress. Hence the cultism of Rorschachers, their allegiances, their bickerings and jealousies, bore and distract him; so also do their endless disputes over—let us say—the criteria of distinction between one F and another. What the clinician wants from the Rorschach worker and researcher are ways of increasing the sensitivity of the test, enlarging its scope, and making it more informative. The present book, on the other hand, leaves the field where it was before it was written.

Perhaps the implied criticism of this review is unjustified in view of the fact that the volume under consideration really purports to do nothing more than reveal "basic processes." If this is so, however, then it must be taken as but another symptom of that compulsive overzealousness to substitute cold hieroglyphs for warm humans that so diverts the stream of clinical endeavor in the behavior sciences.

As for the volume itself, from the point of view of sheer book-making it is a satisfactory job of work. While the material could have been arranged in a manner more considerate of the reader and more sparing on his patience, it must be considered that merely setting up (much less proof reading) the text was a staggering accomplishment.

For those who are preoccupied with scoring problems in Rorschach work and who adhere to the Beck method as exposited already in his *Manual* and other published works, this volume is an amplification and extension of what they already know. But for those who are seeking help in their efforts to help others, this is merely another book bound in blue.

ROBERT M. LINDNER.

Baltimore, Maryland.

WOLBERG, LEWIS R. *Hypnoanalysis.* New York: Grune & Stratton, 1945. Pp. xviii+342.

The hypnosis dealt with in this sober yet provocative work is by and large experimental hypnosis applied to the therapeutic situation, rather than the hypnosis of a purely suggestive treatment. The method includes slow induction of a deep trance, a supervening uninterrupted sleep for about 15 minutes, an hypnotic period comparable in activity to the waking state, during which one or more of the following procedures may be used: Free association to order, dreams to order, interpretation of his own dreams and of other behavior, age-level regression, experimental neurosis, automatic writing, crystal gazing, graphical representation of conflict material, play therapy, dramatic acting, giving of post-hypnotic suggestions, and finally a restful period of uninterrupted sleep. This hypnotic activity seeks not only to discover the traumatic episodes ("enucleation" of them) and the steps whereby false attitudes toward reality (hallucinations and delusions) have arisen

from the patient's projection of his inner frustrations and needs ("empounded self-assertion"), but also to give practice in self-evaluation and even in self-assertion.

Because in hypnoanalysis transference with the stronger personality is indigenous, resistance can be understood and overcome. Since the unconscious strivings and attitudes are nearer to hand, the reality sense can be tested, and all in all, development of the ego can be actively fostered, it becomes an important adjunct to psychoanalysis, or as Dr. Wolberg declares, "Indeed, hypnoanalysis *is* psychoanalysis, performed in a controlled setting." Really, hypnoanalysis is a catalyst by means of which the psychoanalytic dynamisms precipitate out a cure in a short analysis.

Granted a skilled worker, the main limitations of hypnoanalysis are that not all patients can be hypnotized and not all hypnotizable persons have the ego strength to tolerate the transference. The fact that a patient may act out his inner impulses but fail to verbalize them constitutes a hazard (278).

In criticism of this excellent treatise the reviewer would call attention to two points. The first is that hypnosis, as described, depends on what the hypnotist makes of it and, no less, on how the subject reacts to it on the basis of his whole range of interpersonal attitudes (277). It is passive or active, enslaving or liberating, as the two participants consciously or unconsciously wish it—or it may be a stalemate.

The second point is that no new experimental proof is offered for the actuality of age-level regression. Negative findings, such as those reported in 1940 by the reviewer, are brushed aside with a reference to the consensus of opinion, with descriptions of cases, and with a long quotation from Erickson and Kubie describing their second stage of regression as so real that "the hypnotist and the hypnotic situation, as well as many other things, become anachronisms and nonexistent." Certainly, the existence of such a self-consuming stage must be verified under the best of experimental controls.

PAUL C. YOUNG.

Louisiana State University.

TREDGOLD, A. F. *Manual of psychological medicine for practitioners and students* (2nd Ed.). Baltimore: Williams & Wilkins, 1945. Pp. xi +308.

Psychological medicine is concerned with disease of the mind as apart from body. Mind is manifested as conation, affect, and cognition. The first two provide the activating force of conduct. Conduct abnormality is classified as disorder, decay, and defect, each having clinical varieties. Disorder and decay may be primary (functional) or secondary (organic). Defect may be primary (germinal) or secondary (environmental). Disorders arise when an inherited disposition is sub-

jected to environmental stress; in decay inherent defective durability is the principal causal factor. The major portion of the book is devoted to the incidence, symptomatology, etiology, prognosis, and treatment of the clinical varieties. In a general discussion of treatment, psychotherapy is described as explanation, persuasion, and suggestion. There is an adequate index, a 52-item bibliography, but there are no illustrations.

The principal weakness of the book is its outmoded dualism and its consequent perpetuation of a misconception of the functional psychoses as presumably involving no physiological function. The discussion of medical treatments is generally up-to-date but there is no mention of refrigerants or of diet control in epilepsy. Nearly 65 per cent of the bibliographical references antedate 1930.

The first edition (1943), written to fill the wartime need of British medical officers, was exhausted in nine months, hence this edition, which differs solely in a small number of additions, has appeared. Its brief descriptions of the rarer psychoses will make it useful as a reference. However, other adequate sources free from Tredgold's dualistic approach are already available.

STANLEY S. MARZOLF.

Illinois State Normal University.

KVARACEUS, W. C. *Juvenile delinquency and the school.* Yonkers: World Book, 1945. Pp. x+337.

Whether or not juvenile delinquency is a seriously increasing problem cannot be clearly answered with available data; but that juvenile delinquency has been for many years a serious social problem cannot be denied. In the final analysis it is only study of delinquents as individuals which can lead to understanding and perhaps ultimate solution of the problems. The volume being reviewed is the latest in a long series of studies starting with Healy's *The Individual Delinquent* in 1915 which present basic data on the psychological, social, and other characteristics of delinquents determined from studies of individuals.

In 1937 the Children's Bureau was started under the general direction of the Passaic, N. J. Board of Education. Its interest is in truancy and delinquency with the purpose "to eliminate the causes of these offenses, to prevent their occurrence, and to make desirable adjustments for children who have offended." Dr. Kvaraceus, who has been Director of the Bureau, has analyzed the data on the first 761 cases and presents reasonably exhaustive findings in such areas as early development, family, economic status, the community, and the school. In general his findings are in agreement with earlier studies but the group is somewhat unusual in that it contains a relatively high proportion of pre-delinquents.

The recurrent emphasis in this book concerns the place of the school

in the problem of delinquency. The author has no hesitancy in pointing out that frustrating experiences in schools have etiologic significance in delinquency. But he argues well that community programs of dealing with the problem have much to gain if the schools will accept leadership. Certainly the work here described, while not perfect, gives sufficient evidence of success to warrant close study, and imitation, by other communities.

While this work is not particularly oriented by psychologists it will be of value to all those dealing with the delinquent. Its greatest value is as a stimulus to community thinking on a community problem.

C. M. LOUTTIT.

Ohio State University.

BRANDT, HERMAN F. *The psychology of seeing.* New York: The Philosophical Library, 1945. Pp. xvi+240.

This is a difficult book to assess adequately as a unit, because it is essentially a collection of experimental articles loosely tied together with somewhat casual generalizations. When considered separately, many of the individual studies have considerable merit. All of them are based upon a single piece of apparatus, a special camera, invented by the author, which photographs simultaneously the horizontal and the vertical movements of the eyes separately on the same film at points a fixed distance apart. To discover how this is accomplished one must turn to an earlier article listed in the bibliography but not referred to in the text. Illustrations are given, however, of the final projection of the records so that a picture of the excursions of the eyes over the test object is obtained.

There follows a discussion of *Basic Eye Movement Tendencies* with no reference in the text to the rather considerable literature on eye movement. Appending a bibliography hardly compensates for this lack. The reason may be found, perhaps, in the preparation of the original articles for non-technical and semi-popular readers.

Ocular photography is applied, then, to problems of advertising which encompass attention, interest, and preferences; to problems of learning, of art and of perception. Finally, a program of projected studies is laid down, ranging from the measurement of intellectual abilities and aptitudes through determination of the effects of fatigue to detection of guilt or innocence of criminals. The applicability of ocular photography is boundless.

Presentation of the various studies tends to follow a pattern, viz., problem, apparatus, procedure and results. Generalizations and interpretations are made frequently but do not always follow from the data. The attempt to consolidate the various units into a general psychological treatise seems forced and casual. The author's enthusiasm is for the apparatus and its applicability, and this leads to unnecessary repeti-

tions of procedural descriptions and exclamations of satisfaction with the results. Four separate illustrations of the camera are shown, all from so nearly the same angle that they are essentially identical. Such an external view reveals none of the essential features of the device.

Perhaps the reviewer has approached this book from too professional an angle. A suggestion that this is the case is given in the glossary which defines such terms as area, art, color, distraction, fatigue, learning, retina, sensation, etc. Criticism, however, is disarmed by the author's statement in the preface, "The readers may at times sense a repetition of methods of procedures employed in studies assembled and occasional sweeping statements. The common pattern emerged as a result of an attempt to combine studies published elsewhere, and if over-enthusiasm is manifest it is largely because the writer projected his imagination beyond known facts for the purpose of citing potential fields of investigation."

Nevertheless, the generality of the title is misleading. While the problems reported are primarily psychological, they do not, all together, constitute "Psychology" nor all of "seeing." A more definitive title would give a better impression.

FORREST LEE DIMMICK.

Hobart College,
Geneva, New York.